



Authors as users

A deep log analysis linking demographic and attitudinal data obtained from scholarly authors with their usage of *ScienceDirect*

David Nicholas, Paul Huntington and Hamid R Jamali

August 2006

a CIBER Report
UCL Centre for Publishing



Table of Contents

1. Executive Summary	11
2. Introduction	22
2.1. Aims and objectives.....	22
2.2. Scope and parameters of the study.....	22
3. Methods	23
3.1. Questionnaire survey.....	23
3.2. Logs analysis.....	24
3.3. Characteristics of sample.....	25
4. Results	25
4.1. Viewing behaviour (item views and session analysis).....	27
4.1.1 Type of item viewed	27
4.1.1.1 By subject background.....	29
4.1.1.2 By type of organisation	29
4.1.1.3 By age	30
4.1.1.4 By gender.....	31
4.1.1.5 By occupational status	32
4.1.1.6 By geographical location.....	32
4.1.1.7 By number of articles published.....	33
4.1.2 Type of article viewed	34
4.1.3 Session analysis	35
4.1.3.1 By subject back ground.....	36
4.1.3.2 By type of organisation	37
4.1.3.3 By age	37
4.1.3.4 By gender.....	38
4.1.3.5 By occupational status	39
4.1.3.6 By geographical location.....	39
4.1.4 Length of article viewed (number of pages in a paper)	40
4.1.5 Publication status of article viewed	41
4.1.5.1 By subject	42

4.1.5.2	By type of organisation	43
4.1.5.3	By age	44
4.1.5.4	By gender.....	45
4.1.5.5	By occupational status	45
4.1.5.6	By geographical location	46
4.1.5.7	By number of articles published	48
4.1.6	Publication year of article viewed	48
4.1.6.1	By subject background.....	49
4.1.6.2	By type of organisation	51
4.1.6.3	By age	52
4.1.6.4	By occupational status	53
4.1.6.5	By geographical location	54
4.1.6.6	By number of articles published	55
4.1.6.7	By number of views made in a session (Site penetration)	56
4.1.6.8	By number of searches undertaken	56
4.1.7	Individual journal titles used.....	57
4.1.8	Number of unique journals viewed in a session.	58
4.1.8.1	By subject background.....	58
4.1.8.2	By type of organisation	59
4.1.8.3	By age	59
4.1.8.4	By occupational status	60
4.1.8.5	By geographical location.....	61
4.1.9	Subject of journals viewed	61
4.1.9.1	By age of article item viewed	62
4.1.9.2	By publication status of article viewed	63
4.1.9.3	By subject background.....	63
4.1.9.4	By whether respondents thought the quality of an article was determined by the journal in which it was published.....	64
4.1.9.5	By whether respondents thought it is more important to publish in a prestigious general journal, than a MORE appropriate specialised journal	65
4.1.10	Items viewed in a session (site penetration)	66
4.1.11	Return visits	70
4.1.12	Time spent online	74

4.2. Searching and navigating	77
4.2.1 Number of searches in a session.	77
4.2.2 Search approach adopted	80
4.2.2.1 By subject background.....	81
4.2.2.2 By the type of organisation	81
4.2.2.3 By age	82
4.2.2.4 By gender	83
4.2.2.5 By occupational status	83
4.2.2.6 By geographical location.....	84
4.2.3 Gateways	85
4.2.4 Number of returned hits (grouped)	89
4.2.4.1 By subject background.....	91
4.2.4.2 By type of organisation	92
4.2.4.3 By age	92
4.2.4.4 By occupational status	93
4.2.4.5 By geographical location.....	94
4.2.4.6 By age of article item viewed	94
4.3. Attitudinal data.....	95
4.3.1 Core functions	95
4.3.1.1 CERTIFICATION	95
4.3.1.1.1 Where to publish?.....	95
4.3.1.1.2 Peer review/article status	96
4.3.1.2 DISSEMINATION.....	99
4.3.1.2.1 Browsing journals from home	99
4.3.1.2.2 Importance of the digital journal to the user.....	100
4.3.1.2.3 Author's websites as a source of articles	101
4.3.1.2.4 Researcher's raw data.....	102
4.3.1.2.5 Citing behaviour.....	105
4.3.1.3 ARCHIVING.....	106
4.3.1.3.1 Article age	106
4.3.1.4 FUNDING	107
4.3.1.4.1 Difficulty researching new topics	107

4.3.2	Scholarly information seeking behaviour models	107
4.3.2.1	Online Behaviour – Model 1	107
4.3.2.2	Online Behaviour – Model 2	109
4.3.2.3	Online Behaviour - Model 1 informed by questionnaire results.....	110
4.3.2.3.1	Factor 1: Information collectors	110
4.3.2.3.2	Factor 2 – Browsers	111
4.3.2.3.3	Factor 3 - Updaters.....	112
4.3.2.4	Online Behaviour - Model 2 informed by questionnaire results	112
4.3.3	Returnees	113
5.	General Conclusions.....	114
6.	References.....	117
7.	Appendix 1: Supplementary analyses.....	117
7.1.	Top Ten Journal titles.....	117
7.2.	Article versions analysis	121
8.	Appendix 2: Characteristics of sample.....	127

List of Figures

Figure 1 Percentage breakdown of use by type of item viewed.....	28
Figure 2 Percentage breakdown of type of item viewed by discipline.....	29
Figure 3 Percentage breakdown of type of item viewed by type of organisation.....	30
Figure 4 Percentage frequency distribution of type of item viewed by age of users.....	31
Figure 5 Percentage breakdown of type of item viewed by gender.....	31
Figure 6 Percentage breakdown of type of item viewed by status of users.....	32
Figure 7 Percentage breakdown of type of item viewed by 7 regional groupings.....	33
Figure 8 Percentage breakdown of number of articles ever published by item viewed.....	33
Figure 9 Percentage breakdown of number of articles published in last 12 months by item viewed.....	34
Figure 10 Percentage breakdown of article item.....	35
Figure 11 Percentage breakdown of type of article item viewed.....	35
Figure 12 Percentage breakdown of type of article item viewed by subject.....	36
Figure 13 Average number of articles viewed across subject.....	36
Figure 14 Percentage breakdown of article views by format and occupation status.....	37
Figure 15 Percentage breakdown of article views by format and user's age.....	38
Figure 16 Percentage breakdown of article views by format and gender.....	38
Figure 17 Percentage breakdown of article view by format and user's occupational status.....	39
Figure 18 Percentage breakdown of article views by 9 world regional groupings.....	40
Figure 19 Percentage breakdown of format of article by the length of the article in pages.....	40
Figure 20 Distribution of article use by whether article was in press (AIP) or published article (regular): items viewed.....	41
Figure 21 Percentage breakdown of use of articles in press (AIP) and regular articles by subject: items viewed.....	42
Figure 22 Percentage breakdown of use of articles in press (AIP) and published articles (regular) by subject: sessions conducted.....	43
Figure 23 Percentage breakdown of use of articles in press (AIP) and published articles (regular) by organisational affiliation: items viewed.....	43
Figure 24 Percentage breakdown of use of articles in press (AIP) and published articles (regular) by organisational affiliation: sessions conducted.....	44
Figure 25 Percentage breakdown of use of articles in press (AIP) and published articles (regular) by age: items viewed.....	44

Figure 26 Percentage breakdown of use of articles in press (AIP) and published articles (regular) by age: sessions conducted.....	45
Figure 27 Percentage breakdown of use of articles in press (AIP) and published articles (regular) by occupational status: items viewed.....	46
Figure 28 Percentage breakdown of use of articles in press (AIP) and published articles (regular) by occupational status: sessions conducted.....	46
Figure 29 Percentage breakdown of use of articles in press (AIP) and published articles (regular) by major country and regions groupings: items viewed.....	47
Figure 30 Percentage breakdown of use of articles in press (AIP) and published articles (regular) by country groupings: sessions conducted.....	47
Figure 31 Percentage distribution of the number of papers published in the last year by whether author viewed an AIP or regular article.....	48
Figure 32 Percentage distribution of article views by its date of publication.....	49
Figure 33 Percentage breakdown of use between current, declining and old articles by subject.....	50
Figure 34 Percentage distribution of age of material viewed by subject.....	51
Figure 35 Percentage breakdown of use between current, declining and old articles by organisational affiliation.....	51
Figure 36 Percentage distribution of age of material consulted by organisational affiliation of user.....	52
Figure 37 Percentage breakdown of use between current, declining and old articles by age of user.....	52
Figure 38 Percentage distribution of number of unique journals viewed in a session by user's age grouping.....	53
Figure 39 Percentage breakdown of use between current, declining and old articles by user academic status.....	53
Figure 40 Percentage distribution of number of unique journals viewed in a session by occupational status.....	54
Figure 41 Percentage breakdown of use between current, declining and old articles by region.....	54
Figure 42 Percentage breakdown of number of unique journals viewed in a session by country.....	55
Figure 43 Percentage decay frequency distribution by number of articles published in the last year.....	55
Figure 44 Percentage decay frequency distribution over number of views in a session.....	56
Figure 45 Percentage decay frequency distribution over number of searches conducted.....	57
Figure 46 Percentage distribution of number of unique journals viewed in a session by subject.....	58
Figure 47 Percentage breakdown of number of unique journals viewed in a session by organisational affiliation.....	59
Figure 48 Percentage breakdown of number of unique journals viewed in a session by age.....	60
Figure 49 Percentage distribution of number of unique journals viewed in a session by occupational status.....	60
Figure 50 Percentage breakdown of number of unique journals viewed in a session by country.....	61

Authors as users: a deep log analysis

Figure 51 Frequency distribution of views by main subject category of journal.	62
Figure 52 Percentage breakdown by historical value of item viewed over main subject grouping of journal.	62
Figure 53 Percentage breakdown by publication status of item viewed by main subject grouping of journal.	63
Figure 54 Percentage breakdown of journal subject grouping by user subject group.	64
Figure 55 Average score for the statement “quality of an article is determined by the journal” over main subject grouping of journal.	65
Figure 56 Average score for the statement “more important to publish in a prestigious general journal, than a MORE appropriate specialised journal” by main subject grouping of journal.	66
Figure 57 Percentage breakdown of number of views in a session.	66
Figure 58 Percentage distribution of views in a session by subject.	67
Figure 59 Percentage breakdown of views in a session by organisational affiliation.	68
Figure 60 Percentage breakdown of views in a session by age.	68
Figure 61 Percentage breakdown of views in a session by occupational status.	69
Figure 62 Percentage distribution of views in a session by region.	69
Figure 63 Percentage distribution of number of papers published in the last 12 months by number of views in a session.	70
Figure 64 Percentage distribution of number of visits to the site by subject.	71
Figure 65 Percentage distribution of number of visits to the site by organisational affiliation.	71
Figure 66 Percentage distribution of number of visits to the site by age.	72
Figure 67 Percentage distribution of number of visits to the site by gender.	72
Figure 68 Percentage distribution of number of visits to the site by occupational status.	73
Figure 69 Percentage breakdown of number of visits published in the last 12 months by country.	73
Figure 70 Percentage distribution of the number of articles published in the last 12 months by number of visits to the site.	74
Figure 71 Median (seconds) estimate of page view time by item viewed.	75
Figure 72 Median page view time (seconds) for article item views by number of pages of paper.	76
Figure 73 Median page view time by format of article item viewed.	76
Figure 74 Percentage distribution of searches in a session by user subject grouping.	77
Figure 75 Percentage distribution of searches in a session by organisational affiliation.	78
Figure 76 Percentage distribution of searches in a session by user’s age grouping.	78
Figure 77 Percentage distribution of searches in a session by users’ occupational status.	79
Figure 78 Percentage distribution of searches in a session by users’ regional grouping.	79
Figure 79 Percentage of searches by method of searching.	80

Figure 80 Percentage breakdown of different search methods by subject.	81
Figure 81 Percentage breakdown of different search methods by type of organisational affiliation.	82
Figure 82 The distribution of different search methods by user age (sessions).	82
Figure 83 Percentage breakdown of different search methods by gender.	83
Figure 84 Percentage breakdown of different search methods by occupational status.	84
Figure 85 Percentage breakdown of different search methods by country.	84
Figure 86 Percentage of use by gateway (referrer).	85
Figure 87 Percentage breakdown of use by combination of gateway and subject.	85
Figure 88 Percentage breakdown of use by combination of gateway and organisational affiliation.	86
Figure 89 Percentage breakdown of use by combination of gateway and user age.	86
Figure 90 Percentage breakdown of use by combination of gateway and gender.	87
Figure 91 Percentage breakdown of use by combination of gateway and occupational status.	87
Figure 92 Percentage breakdown of use by combination of gateway and country.	88
Figure 93 Percentage breakdown of the number of visits by combination of session type.	89
Figure 94 Percentage of searches conducted by the number of items (hits) returned as result.	90
Figure 95 Distribution of the number of returned hits by type of search option screen.	91
Figure 96 Breakdown of searches conducted by combination of number of returned hits and subject. .	91
Figure 97 Breakdown of searches conducted by combination of number of returned hits and organisational affiliation.	92
Figure 98 Breakdown of searches conducted by combination of number of returned hits and age of users.	93
Figure 99 Breakdown of searches conducted by combination of number of returned hits and occupational status.	93
Figure 100 Breakdown of searches conducted by combination of number of returned hits and country.	94
Figure 101 Percentage decay frequency distribution over use of various navigational tools in a session.	95
Figure 102 Percentage frequency distribution of responses to the Q "It is more important to publish in a prestigious general journal, than a MORE appropriate specialised journal" by the number of views in a session.	96
Figure 103 Percentage frequency distribution of responses to the Q 'Readers do NOT really need refereed journals' by number of different journals viewed in a session.	97
Figure 104 Percentage frequency distribution of responses to the Q "The quality of an article is determined by the journal within it is published" by the number of different journal published in a session.	97
Figure 105 Percentage frequency distribution of responses to the Q 'Readers do NOT really need refereed journals' by form of access.	98

Authors as users: a deep log analysis

Figure 106 Percentage frequency distribution of responses to the Q “The quality of an article is determined by the journal within it is published” by access.	98
Figure 107 Percentage frequency distribution of responses to the Q “I prefer to do my e-journal browsing at home rather than at work” by the number of requests made in a session.	99
Figure 108 Percentage frequency distribution of responses to the Q “I prefer to do my e-journal browsing at home rather than at work” by age of article(s) viewed in a session.	100
Figure 109 Percentage frequency distribution of responses to the Q “An article will only be read if it is available electronically” by article version read.	101
Figure 110 Percentage frequency distribution of responses to the Q “I always search authors’ own websites for the full article” by number of requests in a session.	101
Figure 111 Percentage frequency distribution of responses to the Q “Having greater access to other researchers’ data would benefit my own research” by the number of searches conducted in a session.	102
Figure 112 Percentage frequency distribution of responses to the Q “I am willing to allow other researchers to access my raw research data” by number of requests made in a session.	103
Figure 113 Percentage frequency distribution of responses to the Q “I am willing to allow other researchers to access my raw research data” by number of searches conducted in a session.	104
Figure 114 Percentage frequency distribution of responses to the Q “I am willing to allow other researchers to access my raw research data” by the number of journals viewed in a session.	104
Figure 115 Percentage frequency distribution of responses to the Q “Authors often cite papers when they have only read the abstract” by number of searches conducted.	105
Figure 116 Type of article and abstract view across responses to the question 'Authors often cite papers when they have only read the abstract'.....	106
Figure 117 'It is NOT important to have access to research articles that were published more than 10 years ago' by age of an article viewed.....	107
Figure 118 Percentage breakdown by type of view over top ten (by usage) - journal title.....	118
Figure 119 Percentage breakdown by historical value of item over top ten journals.....	119
Figure 120 Percentage breakdown by article version over top ten journals.....	119
Figure 121 Average score of “quality of an article is determined by the journal” by top ten journals.	120
Figure 122 Average score of “more important to publish in a prestigious general journal, than a MORE appropriate specialised journal” over main top ten journals.	121
Figure 123 Percentage of respondents saying yes to each of the options with regard to which article version is most important for research.	122
Figure 124 Most important article versions for research, grouped into three categories (%).	123
Figure 125 Most important article versions for research, grouped into three categories: by age (%). ..	123
Figure 126 Most important article versions for research, grouped into three categories: by occupational status (%).	124

Figure 127 Most important article versions for research, grouped into three categories: by age of article consulted (%)..... 125

Figure 128 Percentage breakdown of respondents by subject..... 128

Figure 129 Percentage breakdown of respondents by geographical location. 129

List of Tables

Table 1: List of the various item views identified in ScienceDirect logs..... 27

Table 2: Top 20 journals (all article items and including journal menu views)..... 57

Table 3: Factor analysis model 1 based on log metrics to identify academic search behaviour. 108

Table 4: Factor analysis model 2 on log metrics to identify academic search behaviour. 109

Table 5: Principal component analysis rated importance of article type and how respondents deposited their own material..... 126

1. Executive Summary

This substantial (10 page) summary presents the background to the research and its key findings and implications. The body of the report itself contains most of the supporting detail and is presented in such a way that readers can move from the summary to supporting detail with relative ease. The appendices contain a further level of rather more ‘raw’ detail that would be primarily of interest to those whose job requires them to have a deep understanding of the user. In addition mention should be made of a separate document, ‘Authors as users: a subject analysis’ which provides subject portraits of key data presented in this report. While the main report contains a wealth of information never previously having seen the light of day and clearly this is why it needs to be read, its prime purpose and that of the methodology that underpins it is to raise the questions that really need to be asked of the scholarly community. And the report raises many, many questions and its value will come from the answering of these questions.

The main aim of the research project was to provide a comprehensive (360 degree) understanding of the scholarly journal user by linking together data obtained about their attitudes towards scholarly publishing issues and activities with data about their use of ScienceDirect. This has never been done before so the analysis provides unique and rich insights into scholarly journal use. However, because of its novelty, it is inevitably a complex report which deserves some time spent on its contents. A paradigm shift has occurred in scholarly behaviour and this report portrays this on a panoramic scale. In this regard the report provides a synthesis of the data extracted from ScienceDirect usage and search logs and questionnaire responses from 750 authors. Logs were collected for an eighteen month period and in all these authors conducted 16,865 sessions, which saw 110,029 pages viewed.

Simple ‘hit’ or download counts can provide simplistic, vague and misleading estimates of use. Our approach has been to employ six calculations to obtain a robust and comprehensive understanding of viewing and searching behaviour of authors searching ScienceDirect. Together these calculations provide a robust picture of use in the round. The calculations are

- Number of items viewed. An item (screen or page) viewed is defined as the delivery by the server of a single piece of viewable content that has been requested by the client. This item could be an abstract, full-text views, homepage etc. It tells us about the type of content viewed.
- Number of sessions conducted. Users on entering the site for a visit are allocated a session identification number and this defines a session. During a session users might undertake a number of activities, view an abstract, download a full-text, undertake a number of searches etc.

- Number of items viewed in a session. This metric gives the number of items viewed within a user's session and shows depth of user interest or levels of activity.
- Amount of time spent online.
- Number of return visits made a loyalty metric. Users were identified by cookies and the number of times that the user cookie is recorded as returning to the site is taken as estimating the number of return visits.
- Number of individual searches conducted in a session.

ScienceDirect logs record a wide and rich range of information seeking opportunities for the user – searching, browsing and downloading actions, conducted on the entire database or various parts of it. Many of these activities can only be measured by the methodology, deep log analysis, employed in this study and are not a feature of COUNTER compliant data or publisher-produced statistics. Thus the following specialist types of analysis were produced: type of “article” item viewed (HTML, PDF etc); length of article viewed (number of pages); publication status of article viewed (regular, pre-print); publication year/age of article item viewed; individual journal titles used; the number of unique journals viewed in a session; subject of journals used; number of searches in a session; type of search approach adopted (e.g. search engine); number of returned hits in response to a search; differences in usage patterns by method of searching.

Usage was related to user characteristics. For many of these usage analyses the data were further analysed by the following characteristics obtained from questionnaires users filled in: subject background; organisational affiliation; age; gender; occupational status (student, professor, researcher etc); geographical location of the user; productivity (number of articles published).

Attitudes towards various key statements regarding scholarly communication presented in the questionnaire were cross referenced with select usage metrics to yield data about two of the core functions of the scholarly journal – certification and dissemination. This was in order to discover whether usage or information seeking behaviour could be explained by user attitudes or put another way whether attitudes shaped various forms of behaviour. Finally, viewing, searching, demographic and attitudinal data were combined to provide models of scholarly searching behaviour and to relate these to the core functions of scholarly publishing.

The key user behaviour findings were:

1. *Author diversity*. There were very real differences between various types of author, especially in regard to their subject field; academic status and geographical location. This highlights the great danger of trying to generalise data on the back of hundreds of thousands of hits and points to the need to be able to break data down

into discrete communities, which is what the deep log methodology delivers. In other words moving from hits to users. Four findings provide an illustration of this diversity. Firstly, users based in the Social Sciences (31%), Engineering (27%) and Mathematics (23%) recorded high views to articles in print (AIP) and this is clearly an indicator of the importance of currency in these fields. Economics (7%), Medicine (12%) and Life Science (9%) showed lower views and this might be thought to be surprising, especially in the cases of medicine. Secondly, users from Economics (71%), Engineering (71%), the Social Sciences (69%) and Computer Science (70%) made above expected views to articles only one year old. It appears that users associated with these subjects were particularly interested in current material. Thirdly, Material science (39%), Mathematics (38%), Physics (33%) and Chemistry (32%) users were most likely to view 2 or more journals in a session. Fourthly, not surprisingly, users from a subject tended to use the journals related to that subject. However, there were real differences, indicating, perhaps, differing levels of interdisciplinarity. Thus, 71% of those describing themselves as physicists viewed physics journals and this subject registered the highest accord between user discipline and journal discipline (in other words it was the most self-contained). However, users from the Environmental (34%) and Computer Sciences (34%) were least likely just to view journals within their discipline. Further examples of author diversity follow.

2. *Returning users.* Forty percent of ScienceDirect users just visited once over the 5 month survey period, 24% visited 2 to 5 times, 15% visited 6 to 15 and 21% of visited over 15 times. Typically, elsewhere we have found that a high proportion of users are what we call 'bouncers' - people who bounce into a site from a search engine, don't do much and then don't come back. In information terms they are promiscuous. However, the number of authors coming back to search ScienceDirect was greater than we have found elsewhere and points to the fact that: a) scientists are generally more frequent visitors because of the nature of their subject field (ScienceDirect is more scientific in content than the other services - Synergy, OhioLINK, we have studied; b) ScienceDirect has a more loyal body of users, due to the popularity of the titles it offers, the kind of people that use it and the general quality of the product; c) authors are more loyal users than users in general. Physics (81%) and Computer sciences (80%) recorded very high percentages of repeat visits (these subjects also recorded high percentage of current article users). These are very loyal and 'needy' users indeed. The likelihood that a user will repeat their visit increases with age. Women were generally more likely to return to the site compared to men; 66% visited 2 or more times compared to 58% for men. Those people visiting more regularly tended to publish more, which confirms what we

might have expected. About half of those visiting 6 or more times had published five or more papers in the past year compared to a third of those who had done so who had just visited once.

- *Site penetration (number of views in a session)*. This is a ‘busyness’ indicator. Typically users of e-journal libraries do not view many items in a session – ‘quickly in and out’ describes nicely this very consumer form of behaviour. However, ScienceDirect users not only returned more often as we learnt above, but they also penetrated the site more deeply – again this may be explained by the fact they were authors as well. Thus over a third of users viewed 4 to 10 items (articles, abstracts, searches etc.) and 16% more than 10 items in a session. The number of viewed items tended to increase with age and combine this with the fact that so does the number of visits suggests age is a very significant factor in explaining scholarly information seeking behaviour. Asian and South American users recorded with the greatest number of views and African and Eastern European users the least. Of the individual countries Germany and China have particular active users. Analysing the data by geographical region or country demonstrates quite conclusively that, while science might be global, usage certainly differs enormously according to country and region. In terms of academic status, students were a third more likely to view just 1 to 2 items in a session compared to senior research staff – no doubt, checking out references given to them. Similarly those under 36 were about a 33% more likely to view just 1 to 2 items compared to those aged 36 to 45. There were sizeable gender differences as well, with men being about a quarter more likely to view 1 to 2 items in a session, as compared to women. Those entering via a gateway link were about 5 times more likely to just view 1 to 2 items in a session as compared to those accessing the site other than gateway – possibly have done all their searching in another database, like PubMed and just come in to ScienceDirect for the full-text. Users viewing most items in a session were more likely to have published articles in the last 12 months: 60% of those viewing 11 or more items had published 5 or more papers, compared to about 50% of those viewing 3 or less items. This shows a strong relationship between production and use.
3. *Length of an article*. A much more specific finding, but possibly very significant, relates to the length of an article viewed. The greatest amount of average online time was spent by users on papers 4 to 10 pages long and the least amount of time on papers 21 pages or more in length. This suggests that people spend more time reading shorter articles online. This might mean that shorter articles are more likely to be read than longer ones, and it follows maybe, more likely to be cited. Many of the articles downloaded, to be read at another time, may never see the light of day again. A related finding supports this supposition: as the size of a paper (measured

by the number of number of pages) increased there was an increased likelihood that the item was viewed as an abstract or as a summary plus and there was less likelihood that the item will be viewed in a PDF or Full text format. In this regard the so-called gold standard metric of downloads needs questioning and investigating and possibly also whether a significant (and possibly worrying) change of user behaviour is occurring. Nowhere in the published literature have we seen this mentioned.

4. *Abstracts.* Abstracts were once thought to redundant given the widespread availability of full-text but they are very much valued by the user as a means of speed reading, relevance checking etc as this report demonstrates. However there are marked variations in their take-up. The use of abstracts, except for the age group 26 to 35, tended to increase markedly with the age of the author. There was a tendency for a greater number of researcher sessions to just view abstracts or summary plus items, perhaps a case of researchers ranging widely in their pursuit of information. Social Scientists conducted the highest frequency of sessions just viewing abstracts items (41%). Senior researchers were more likely to view just abstracts 25% did so, but this group might be under the greatest time pressure. Users from Australasia and Eastern Europeans were proportionately big conductors of abstract only sessions. Possibly, they are non-subscribers?
5. *Full-text downloads.* For many libraries and publishers this is the 'true' indicator of user satisfaction, although as we have mentioned this needs revisiting. Students made the greatest use of full text (HTML) articles and Chinese users recorded the highest use of PDFs. It is postulated that undergraduates prefer HTML from which they can copy and paste the material into their assignments, while academic staff, prefer a PDF format as this is as close as possible to the look and feel of the published hard copy example.
6. *Articles in press.* Sixteen percent of article views were to articles in press (AIPs) and 84% to finished articles. Authors from the fields of Business & Management, Engineering and Mathematics recorded higher than expected views to articles in press. However, it should be recognised that AIP published items are not evenly distributed over subject fields and this could explain it and should be a further investigated. The younger the user, the more likely they were to use articles in press. About a quarter (23%) of those aged under 36 viewed articles in press, however, this was only true of 6% for those aged over 65. It is hypothesised that younger researchers are keen to secure a research advantage by being more up to date and hence are more interested in AIPs, author websites, depositories etc. Those people just undertaking articles in-print sessions were less likely to be prolific

authors (i.e. published 5 or more papers). This could be because of author's greater need to see and cite the final product.

7. *Successful searching.* It is possible that high levels of 'hits', which are perceived to be a measure of success, actually mask real problems with the system, and, maybe, that much information seeking is actually negative or problematic. Thus it appeared that about 40% of users either abandoned their search after viewing the search screen or were returned zero matches for the search they entered. Whether this represents 'failure at the terminal' (user confusion) or the volatility of information seeking behaviour, with people rapidly abandoning search strategies in the rapid pursuit of data is not clear, but there are very serious issues that need researching here. Of those sessions where a search was undertaken half saw just one search conducted, 35% saw 2 to 4 searches, 9% 5 to 10 and 1% over 10 searches. The likelihood of undertaking online sessions where only one search was executed increased with age. Over two-thirds of searches by authors aged over 65 recorded zero returns, which suggests that there could be problems with elderly users. Not surprisingly, research staff were most likely to complete 2 or more searches in a session and about 47% did so.
8. *Obsolescence/decay.* We have found elsewhere that databases and search engines have increased the visibility of older material and this has led to an increase in their use. However, the picture is not quite so conclusive in regard to ScienceDirect usage. Thus well over half (58%) of the article items viewed were to the current one-year period, a third (35%) were to articles aged 2- 6 years old and only 7% were to articles 7 years or older. However, we did find that those using a search option, rather than navigating hierarchical menus, to locate material were less likely to view current material and were more likely to view historical material. Those people aged 36 to 55 were also proportionately more likely to view material older than one year. This age group however were more likely to be senior researchers and were thus more likely to conduct deep searches. Those viewing more pages in a session viewed a greater range of historical material and those conducting more searches were also more likely to view older material. Those users viewing a wider range of historical material tended to prolific authors: 60% of those viewing a combination of aged material, including old, had published 5 or more papers and this compares to about 50% who had done so for those users who had just viewed current material in a session. It was older and younger authors, those from Spain and China, Business studies authors, Hospital staff and students who were more likely to view current material.

Authors as users: a deep log analysis

Relating online user behaviour to scholarly activities and attitudes produced valuable additional insights. In general it highlighted the differences between experienced and less experienced authors, the former being generally much more active information seekers and this is an important group publishers should identify and target. More specific findings follow, thus in regard to the following user attitudes to statements pertaining to core scholarly journal functions:

- “It is more important to publish in a prestigious general journal, than a MORE appropriate specialised journal”, it was found that those viewing more journals were more likely to agree with this statement. These users tend to be the more experienced authors.
- “I prefer to do my e-journal browsing at home rather than at work”, it was found that those: a) viewing more items in a session were more likely to disagree with the statement – the explanation for this could be that those viewing more items in a session maybe using the university print facilities to print out articles.; b) viewing older material were more likely to disagree with this statement. This is probably explained by the fact that those viewing more items in a session tend to be those that search or navigate to material and they tend to view older material.
- “An article will only be read if it is available electronically”, it was found that: 1) those just viewing articles at the in press stage in a session strongly disagreed with this statement. It is thought this might reflect the views of the young post graduate researcher, these users favour AIPs but prefer to print out, or to have saved, the PDF version that gives them the look and feel of the real journal article. 2) women were more likely to agree with the statement. No real explanation for the latter, so need following-up.
- “I always search authors’ own websites for the full article”, it was found that those respondents: 1) viewing more items in a session were more likely to strongly agree – a case of active searchers being more active information-wise generally; 2) reviewing more papers in the last 12 months were more willing to agree to this statement – another case of active scholars being generally active on all fronts.
- “Informal sources of communication such as conferences, bulletin boards are NOT important in scholarly publishing”, it was found that: 1) those publishing more articles in the last 12 months were more likely to agree – these people were after all heavily involved with the formal system; 2) research staff were likely to strongly agree with the statement – these people, too, were heavily involved with the formal system; 3), not surprisingly younger users (and students) were less likely to agree with the statement.

- “Having greater access to other researchers’ data would benefit my own research”, it was found that those: 1) conducting more searches were more likely to strongly agree with the statement; 2) those completing a greater number of searches were less willing to share their research data – more competitive people, perhaps; 3) viewing more journals were less likely to agree to this statement;
- “Authors often cite papers when they have only read the abstract”, it was found that those respondents performing more searches were less likely to agree with this statement. Again a link that need further investigation.

What relating usage to respondent’s views (as expressed in a questionnaire) also delivered was a kind of ‘triangulation’, which enabled us to determine whether what was actually done fitted with what was being said. At one level it enabled us to determine the validity of questionnaire responses. Two analyses conducted are particularly illustrative in showing what can be shown:

- Analysis examining importance of peer review/article status. In this connection we sought to determine whether respondents who agreed with the statement 'Readers do NOT really need refereed journals' searched differently from those that did not, with the hypothesis being that those people who disagreed would limit their searching to a more selective group of titles. It was found that the hypothesis was true with those people viewing a lot of journals in a session agreeing that readers did not really need refereed journals.
- Analysis examining age of material consulted. This analysis examined the relationship between responses to the question 'It is NOT important to have access to research articles that were published more than 10 years ago' to the age of the articles viewed. As might have been expected those who strongly agreed with the statement were most likely to engage in a session where just current material was viewed and those that disagreed were most likely to engage in a session where only older material was viewed. Powerful triangulation here.

Two scholarly information seeking behaviour models were developed using factor analysis; the first based purely on aggregated data from log files and the second, which considered whether an event happened or not within a session. For example, in the first model the number of searches conducted by each user is considered, while in the second we were only concerned if the user did a search or not. Two models were generated as the data presented difficulties with regard to extracting factors. Neither model is better or worse but both should be considered as providing additional views on to the data.

Model 1 identified three main behavioural groups of authors: Information collectors, Browsers and Updaters.

Authors as users: a deep log analysis

Information collectors: Users evidencing this kind of behaviour examined a number of different journals in a session, viewed declining material and, to a lesser extent, current and old material. They were also likely to look at a number of abstracts and favoured viewing in full text mode. These people ranged around widely for information. Their key demographic characteristics were that they were more likely to come from: a) Material Science and Mathematics; b) Germany, Netherlands and Canada – and were less likely to come from China.

In regard to their attitudes Information collectors were 1) likely to strongly agree with the statement that they “search authors' own websites for the full article”; 2) less likely to search from home; 3) more likely to agree that they “published to secure funding/tenure”. 4) more likely to agree with the statement that “peer review does NOT improve article quality”; and, 5) less likely to know about 'institutional repositories'

Browsers: Users evidencing this form of information seeking behaviour obtained information by just viewing journal homepages and journal issues. Their key demographic and attitudinal characteristics were:

- they tended to come from Business Management, Chemical Engineering, Chemistry, Environmental Science and Mathematics ;
- they tended to be younger and male;

In regard to their attitudes they were: 1) less likely to agree with the statement that articles will “only be read electronically”; 2) more likely to agree with the statement “conferences, bulletin boards are NOT important in scholarly publishing”; 3) less likely to agree with the statement “The publisher adds little value”; 4) more likely to agree that they “published to secure funding/tenure”; 5) more likely to agree with the statement that “peer review does NOT improve article quality “.

Updaters: In the case of updating behaviour users viewed current material, were unlikely to view declining material and viewed articles in print. They could be thought of as active users. These users viewed full text content in PDF. Their key demographic and attitudinal characteristics were:

- They were less likely to come from Chemical Engineering and Mathematics;
- They tended to be older (65 & over) and female;.
- Post graduates favoured this form of information seeking;

As regards attitudes, Updaters were: 1) more likely to agree with the statement that they “search authors' own websites for the full article”; 2) more likely to agree with the statement “Quality of an article is determined by the journal”; 3) more likely to agree that they published

to secure funding/Tenure; 4) less likely to agree with the statement that “peer review does NOT improve article quality” 5) more likely to know about 'institutional repositories'

Model 2 generated five main behavioural groups: Gateway users, Searchers, Search/browsers, Abstract viewers and Browsers.

Gateway users: These users came in via a gateway, did not use hierarchical menus (journal issues), on average preferred current material (perhaps determined by their method of access), would not generally view articles in press and tended to view both PDF and Full text documents. Their key attitudinal characteristics were they: 1) more likely to believe in the peer review process; 2) more likely to agree that the “publisher adds little value to the article”; 3) more likely to agree that they published to secure funding/tenure; 4) more likely to agree that authors will pay to have their articles published

Searchers: These users tended to use the search facility, were average users of declining material, below average users of current material and tended to view a number of different journals. These users viewed items in either PDF or Full Text, but not both. The Searchers key attitudinal characteristics were they: less likely to agree that they published to secure funding/tenure; less likely to agree that authors will pay to have their articles published; less likely to agree with the statement that they were unable to review the literature as thoroughly due to time constraints; less likely to know about 'institutional repositories'.

Search/browser users: These users tend to use a combination of the search engine and menus - journal issues, they had a tendency to view current material - seemingly in both formats (PDF/Full) and they viewed a number of different journals. These users were less likely to visit just once and were more likely to be regular users.

Abstract viewers: These users predominately viewed abstracts and not PDFs and tended to view current material. Perhaps these users were non subscribers or came in via Google (bouncers?).

Browser: These users were gathering information just using menus - journal issues; perhaps, they have not been able to work out how to use the search facility because they seemed to be avoiding it or they simply knew the parameters in which they wanted to conduct the search (i.e. a particular journal). They were viewing both abstracts and declining material. These users viewed items in either PDF or Full Text but not both.

The combination of logs and questionnaire data has provided all kinds of interesting results regarding authors and their scholarly practices and views. It has also raised questions that we have attempted to answer but really for full understanding we need raise these questions in the next author questionnaire. This was a pilot study which has demonstrated profitable and (a few) unprofitable lines of investigation and raised questions that urgently need answering. Some of the data are particularly challenging in terms of interpretation and

Authors as users: a deep log analysis

consideration of how the attitudinal data can be best employed is a particular example and in future it would be best to drive the questions off the log data rather than simply look for association. With the knowledge we have acquired and a methodology in place now for investigations of this power, a similar study on a bigger sample population, say, 3000 authors should be undertaken. ScienceDirect has a huge lead over its competition in understanding the user and this would ensure it maintained its lead.

2. Introduction

CIBER at UCL are an independent information science research group. As a result of four years of user research in fields as varied as newspapers, health and scholarly publishing, we have developed what we believe to be a unique set of methodologies (deep log analysis), which provides the most efficient and effective method for monitoring the use and impact of global digital online services, like ScienceDirect. This methodology is essentially based on the analysis of raw transactional server log files (digital fingerprints) and the relating of these to datasets containing user information. We have conducted pioneering deep log work with Emerald, IoP, Blackwell and most recently with OhioLINK and OUP, but we believe this is the very first time that raw usage (and search) logs have been related to user data in the form of online questionnaire returns to provide a robust and 360 degree view of the user. These data have, of course, been anonymized but nevertheless provide kinds of analysis not seen before.

The research was conducted independently from ScienceDirect team and the direction of the research and its analysis were wholly the responsibility of the research team.

2.1. Aims and objectives

The general aim of the research project was to provide a detailed, insightful and independent deep log evaluation of authors who were also users of ScienceDirect. This was to be achieved by linking an online questionnaire distributed to its authors, which provided demographic and attitudinal data, with raw transactional server log data generated by the same people when they searched for, and viewed, material on the ScienceDirect website. Authors are clearly a strategic group about whom we know very little.

This would not only provide a comprehensive understanding of ScienceDirect users and their information seeking behaviour, but also, importantly, add to our understanding of some of the core functions of scholarly publishing (Mabe and Amin, 2002). The ones investigated in this report were certification, dissemination, archiving and funding.

The secondary aim of the investigation was as much to test an innovative methodology which would help publishers who run major questionnaire studies, relate that data in a meaningful way with the vast mountains of log data they accrue automatically on a daily basis. In methodological terms this helps deliver triangulated data: transactional log (limited record of actual population behaviour); questionnaire data (self reported sample behaviour); and observation (small sample fixed task laboratory behaviour), which is not covered in this report.

2.2. Scope and parameters of the study

The study was undertaken in regard to ScienceDirect author/users. The study combined three sets of data: usage logs (a record of what people viewed and when), search logs (a record of how people got there, what types of searches they conducted) and a questionnaire which

provided demographic and attitudinal data which could be related to what people did on ScienceDirect to see what the relationships were.

The data sets were:

- **Usage data.** Three types of usage data were generated to provide a robust and comprehensive picture of the: 1) number of items viewed; 2) number of search sessions conducted; 3) amount of time spent online.
- **Search data.** Search approach/options chosen; gateways used; search options adopted; number of returned hits; number of searches in a session.
- **Questionnaire data.** Data was extracted from a questionnaire sent to authors regarding the following: a) demographic characteristics: subject/discipline; type of organisation for which they worked; occupational status (student, professor, researcher etc); paper productivity; age, gender, and geographical location; b) attitudes to a number of scholarly communication issues/behaviour.

ScienceDirect is a comprehensive web-based collection of STM journals, books and bibliographic information. ScienceDirect was launched in beta form in 1997 and its first commercial licensing started in 1998. ScienceDirect now offers access to more than 2,000 journals, more than 7 million full-text articles and more than 75 million abstract records, from all fields of science. Additionally, users gain online access to multimedia features not available in print journals, such as video, audio and spreadsheet files.

3. Methods

The investigation was all about relating questionnaire data to usage/search data as evidenced in the logs thus the first step was to ensure that the data could be linked.

3.1. Questionnaire survey

The questionnaire from which the demographic and attitudinal data was generated was an online questionnaire which went out on May 2005 to 49,266 authors and obtained 6,344 responses. The respondent's IP number was collected as well as their written responses. This database of anonymous IP numbers was then compared to IP numbers collected by the server used to access the ScienceDirect web site over the previous 18 months. Those IP numbers that could be safely associated with a single computer were selected. Thus, in order to derive a unique user, we only extracted data from the transaction log file for those IP addresses that had a unique machine cookie. Machine cookies are dropped on a client computer when a user accesses ScienceDirect. If an IP address in the log file had a number of different machine cookies associated with it then one of the following might be true: more than one user is using the machine and this would cover proxy IP addresses; the client machine is allocated different IP numbers on different occasions as would be the case with those client machines that are part

of a floating IP network; or lastly that the client's machine is deleting cookies either automatically or by the user. For this study only log entries for IP addresses with one machine cookie were used. That is most of the anomalies surrounding non-unique IP addresses as a result of proxy servers and floating IP addresses have been excluded and hence a pretty good mapping of individual user, who filled in the questionnaire, to log entries was obtained. It should be stressed that the questionnaire data was anonymous as was the IP address data, and at no point could an individual be identified.

3.2. Logs analysis

Server log data are records of actual web pages viewed. These records occur as a result of requests made by the clients' computer and provide a record of pages delivered from the web server to the clients' computer. Log entries were tracked back over an 18 month period. The following gives an example of the ScienceDirect log file:

```
134.5.159.61, 143915, fc0f2bc6-b9e5-11d9-975c-8a0c5905aa77143915, 05/01/05,
02:09:57, C000061700, , 298789480, SearchQuick_Search, 2, n, Media_Searched, allinprod
```

The first field (134.5.159.61, 143915) provides the IP address. This is an anonymous machine-to-machine address number used by computers to correctly send and receive data over the internet. The second field (143915) is a _cookie and is used by the server to recognise a machine that has requested information previously. The third field (fc0f2bc6-b9e5-11d9-975c-8a0c5905aa77143915) is a session cookie and is a number the server uses to track transactions within that session. The fourth and fifth field (05/01/05, 02:09:57) provide the date, time and time record of the transaction. The sixth field (C000061700) is the users' account number. No information was supplied that enabled user account details to be linked to the database. The seventh field (blank in this example) records the previous site visited immediately prior to accessing ScienceDirect frequently this will be a gateway such as PubMed etc. The eighth field (SearchQuick_Search) records the event identifier. The ninth field (2) records the functional area descriptor. The tenth field (n) is the session event snr. The eleventh field (Media Searched) records the attribute type name. The twelfth field (allinprod) records the attribute value description.

The important working definitions adopted for the analysis are as follows:

- ✓ **User.** Users are researchers who have published a paper. User identification was based on the IP number and enhanced by cookie information. A user is effectively a computer; sometimes that computer represents an individual, (i.e. a professor in his office); however, the name of the individual remains anonymous.

- ✓ **Sessions.** A search session in which a number of actions are undertaken. They are identified in the logs by a session identification number.
- ✓ **Items viewed/requests made.** A ‘complete’ item returned by the server to the client in response to a user action. Typically this might be an abstract, an article or a table of contents. A complete item might be all the pages, charts etc. from an article and this is recorded as a single item and hence is quite different from traditional server log files that record pictures and text documents separately.

Metrics will be compromised by caching. This occurs when the client views previously requested pages from the cache on their computer. Caching is the storing of previously viewed pages on to the client’s computer; repeat in-session accesses to these pages are made from the cache and are not requested from the web site’s server and hence not recorded in the logs. This impacts on views as a result of backward navigation. Studies into navigation behaviour (Tauscher and Greenberg, 1997) have shown that backward navigation accounts for almost a third of navigation actions. Caching results in the underreports the number of pages viewed, also impacts on number of views in a session. Fieber (1998) estimates that between 35 to 55% of page impressions are not recorded as they have been cached to the client’s local machine.

3.3. Characteristics of sample

In all there were 757 authors who completed questionnaires which could be matched to the ScienceDirect usage logs. These users viewed 110,000 items. See Appendix 2 for details.

4. Results

Deep log analysis can produce an extremely wide range of analyses, many of which can only be produced by this form of analysis. Add in questionnaire data and we have an almost unlimited number of analyses.

Typically, proprietary software just calculates behaviour in relatively simple and crude terms – number of hit or views. However, in the case of the deep log analysis methods employed here, a far more sophisticated approach has been adopted. Five calculations have variously been used throughout the paper to measure viewing and searching behaviour:

- Number of items (screens, pages) viewed. The number of items viewed is a simple and direct metric that is useful in that it provides the big picture in regard to information seeking, it provides aggregated data.
- Number of sessions conducted, number of items viewed in a session (site penetration). A user session included a range of transactions, views, uses etc conducted as part of the visit to the site by a particular user. It is from the session data that we can witness the pattern of user interactions and methods of navigation.
- Amount of time spent online;

- Number of return visits made
- Number of searches conducted in a session (this is also a navigational metric)

The ScienceDirect logs record a wide range of viewing opportunities – searching, browsing and downloading actions, conducted on the entire database or various parts of it. The full list of the item views analysed was:

- type of item viewed (article, search, menu etc)
- type of article item viewed (abstract, PDF etc)
- length of article viewed (number of pages)
- publication status of article viewed (regular, articles in print)
- publication year/age of article item viewed
- individual journal titles and subject used
- the number of unique journals viewed in a session,
- subject of journals used

The ScienceDirect logs disclosed how users searched or navigated their way to and around the database. The following analyses were conducted:

- number of searches in a session
- search approach adopted
- number of returned hits;
- differences in usage patterns by method of searching;

For many of the above analyses the data were further analysed by the following user (respondent) characteristics extracted from the questionnaire:

- subject background
- organisational affiliation
- age
- gender
- occupational status (student, professor, researcher etc)
- geographical location of the user
- productivity (number of articles published)

Respondent attitudes towards various key statements regarding core scholarly functions presented in the questionnaire were cross referenced with select usage metrics. The statements were:

- The quality of an article is determined by the journal within it is published (Certification)
- I prefer to do my e-journal browsing at home rather than at work (Dissemination)

Authors as users: a deep log analysis

- An article will only be read if it is available electronically (Dissemination)
- It is more important to publish in a prestigious general journal, than a MORE appropriate specialized journal (Certification)
- Having greater access to other researchers' data would benefit my own research (Dissemination)
- I am willing to allow other researchers to access my raw research data (Dissemination)
- Authors often cite papers when they have only read the abstract (Dissemination)
- It is becoming increasingly difficult to carry out research in new and interesting areas (Funding)
- It is NOT important to have access to research articles that were published more than 10 years ago (Archiving)

Finally, viewing, searching, demographic and attitudinal data were combined to provide models of scholarly searching behaviour and to relate these to the four core functions of scholarly publishing.

The following results section is divided in to three subsections: 4.1 viewing behaviour; 4.2 navigational and searching behaviour; 4.3 attitudinal data seen in the context of viewing behaviour

4.1. Viewing behaviour (item views and session analysis)

In total the logs of 757 user-authors were analysed. These users conducted 16,865 sessions, which saw 110,029 items viewed over a period of eighteen months.

4.1.1 Type of item viewed

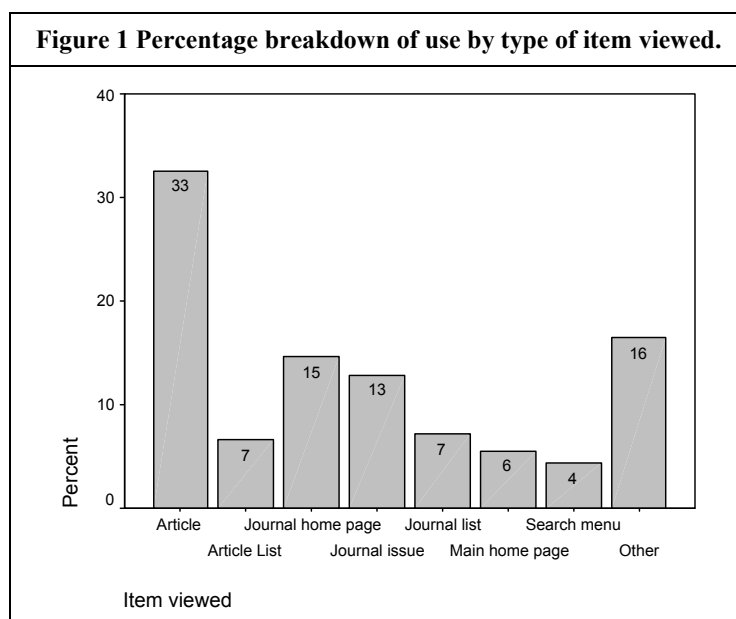
An explanation of what the categories means is necessary to the understanding of this analysis and all item view including access figures. This can be found in Table 1.

Table 1: List of the various item views identified in ScienceDirect logs.

Article	Journal, book series or handbook series article in any format (html, PDF, abstract, summary plus - an enhanced abstract). Includes images and references with hyperlinks
Article list	Table of contents of an individual journal issue
Gateway	User searches ScienceDirect via a third party site rather than use menus and search facilities available on ScienceDirect. Links directly at the article level on SD
Journal home page	The home page of a journal
Journal issue	The issue page of a journal
Journal list	A list of journals
Main home page	Home page of ScienceDirect.

SearchMain	ScienceDirect's search facility. There are 7 options on the homepage, one links to main search page; the others are links to home, journals, books, abstract database, my profile, and alerts. This is the screen that opens after clicking the Search button. All available tabs except for the Scirus tab are included in this functional area
Others	<p>Thisjrn. When viewing an issue, table of contents or abstract page of a journal, one of the search options is to search this journal and that restricts the search to the given journal.</p> <p>Media Search is an attribute of Quick Search. It tells you which part of the product is searched. The values available in the dropdown on SD depend on the context. E.g., if you are on a journal page you can search that journal, all journal or all full-text searches. If you go down to the issue level you also have the option to do a Quick Search within the volume/issue.</p> <p>Journal. A search limited to the Journal area.</p> <p>Allsource This option (on the search page) is a search of all of the sources including journals, books, and the abstract databases. Search in all of the journals of ScienceDirect</p> <p>Alljnl. A search through the whole SD product</p>

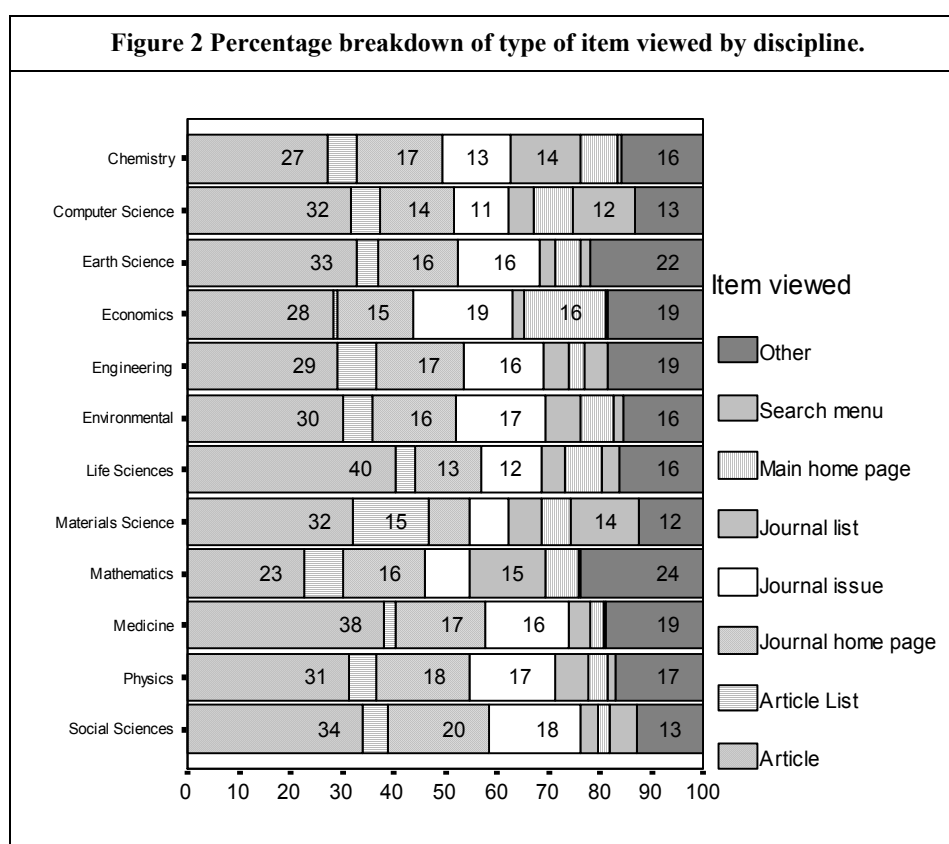
Figure 1 reports the percentage frequency distribution by type of item viewed including access. Articles (for different formats, see below) were the most viewed type of item and accounted for just under a third, 33%, of all views. The second most important item was the journal home page (15% of views) followed by Journal issue pages (13%). This finding does not present a complete picture of downloads due to the aforementioned problems of caching and the searching of title and bibliographic information on other Gateway sites.



Authors as users: a deep log analysis

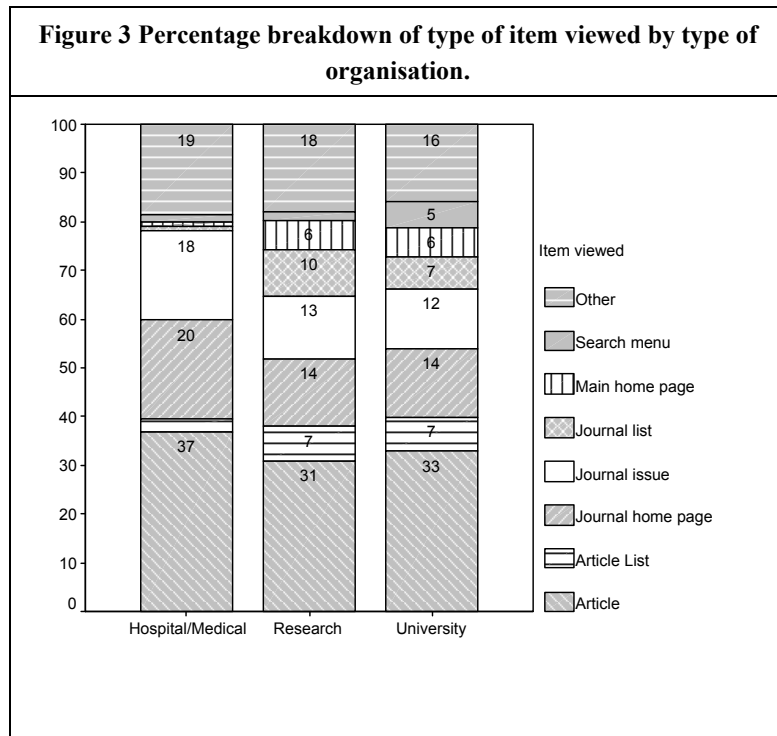
4.1.1.1 By subject background

Figure 2 gives the percentage frequency distribution of type of item viewed by discipline. The profiles were quite different in the case of users from Social Sciences and Physics. Respondents from Social Science (20%) and Physics (18%) made above expected use of the Journal home page. Life Sciences (40%) and Social Sciences (34%) downloaded proportionally more articles than Mathematics (23%), Chemistry (27%) or Economics (28%). Material science (14%) and Computer Science (12%) respondents were heavy users of the search facility. The journal is seemingly more important to Social Sciences and Physics as the use of the home page is higher than that for other groups



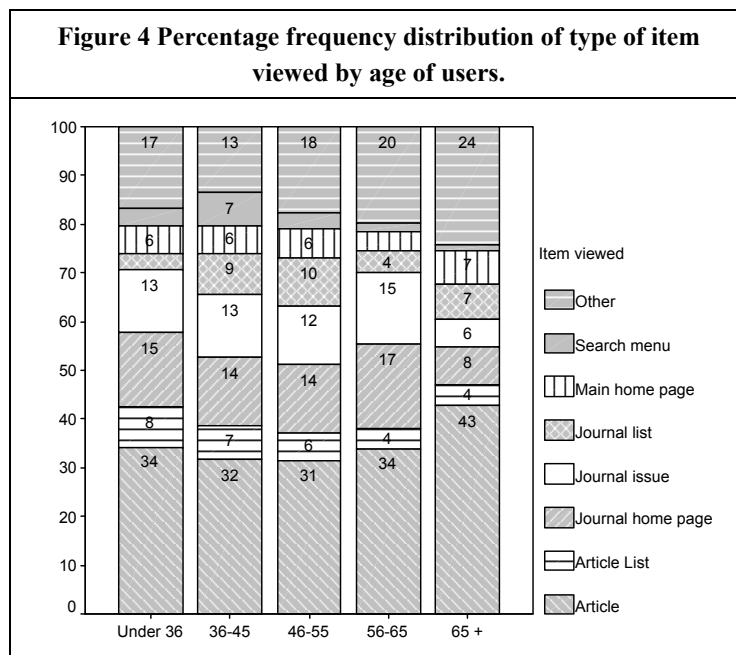
4.1.1.2 By type of organisation

In terms of organisational affiliation those respondents working in hospitals (37%) viewed relatively more full-text articles than the other groupings; furthermore, hospital users were also above average users of the journal home page (Figure 3). This finding requires explanation in follow-up surveys



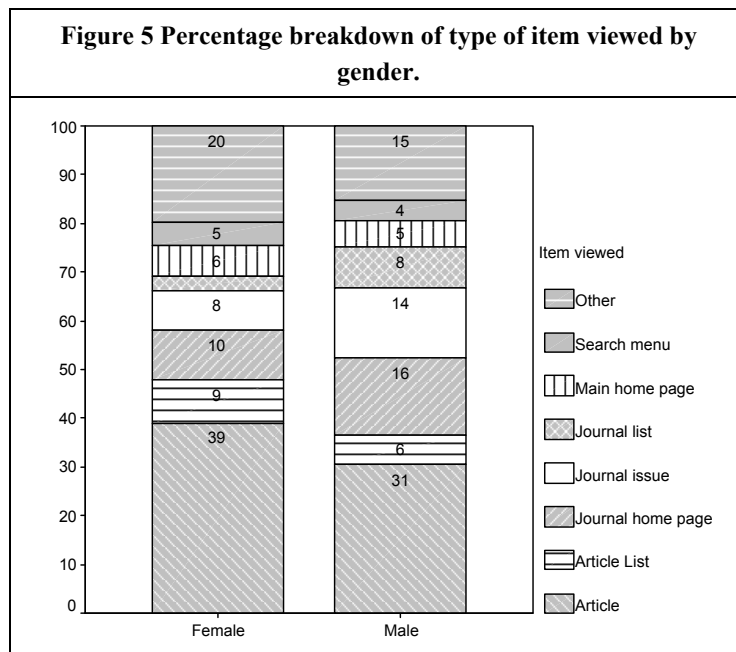
4.1.1.3 By age

Those aged 36 to 55 tended to be the heavy users of the journal list, which accounted for 98% (36 to 45) and 10% (46 to 55) of views (Figure 4). Those aged 36 to 45 used the search facility more (7% of views). The use of the Journal home page varied with age from about 14% for the age group 36 to 45 to 17% for those aged 56 to 55. Use of the journal list declined with age from about 8% for the age group 36 to 45 to 4% for those aged 56 to 55. It is thought that as users age they develop time saving methods to access material and this generally results in less hierarchical menus being viewed. They are also more able to judge a journal by its merits.



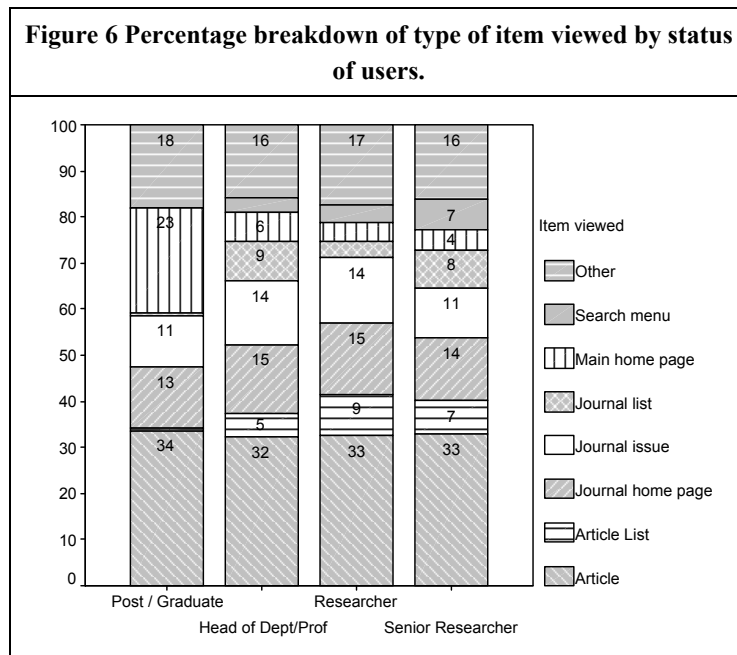
4.1.1.4 By gender

Interestingly, women tended to view more articles than men, 39% did so, compared to 31% of men (Figure 5). Men made more visits to the journal home page, 16% compared to 10% and greater use of journal issues - 14% compared to 8% and journal lists 8% compared to 2%. Quite significant differences here but an immediate explanation for this is not obvious.



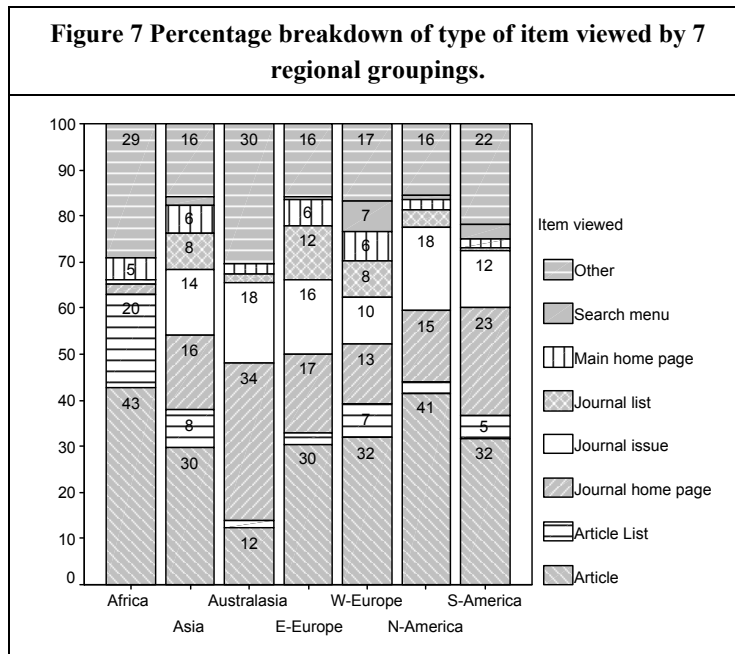
4.1.1.5 *By occupational status*

Issue article lists were used most by researchers (9% of views) and senior researchers (7%). Senior researchers (8%) and Heads of Department (8%) made greater use of Journal lists as compared to researchers (3%) and graduates (0%). Senior researchers made greater use of the search facility, 7% did so (Figure 6).



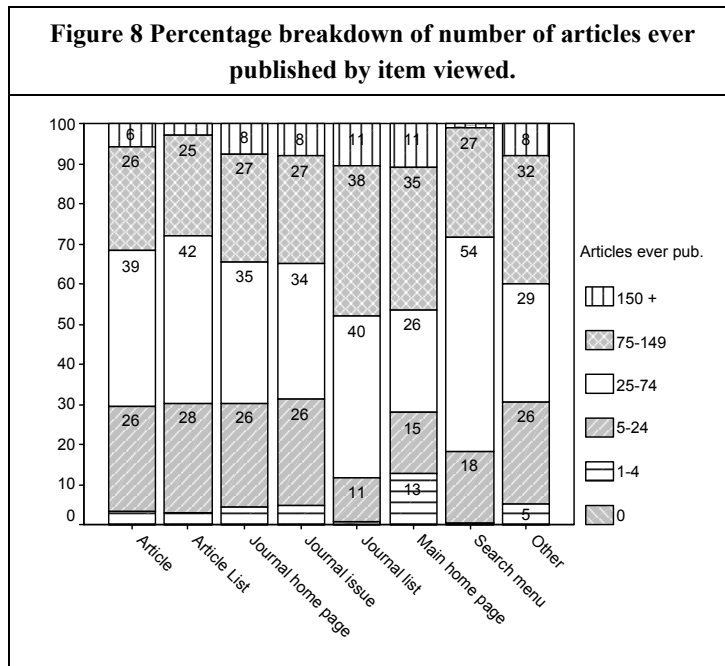
4.1.1.6 *By geographical location*

In terms of the geographical location of respondent (Figure 7), there were some quite big differences between the use profiles of the regions. North Americans were relatively low users of the search facility (3%), while those from the Western Europe (7%) were heavy users. Users from Australasia tended to view the journal home page (34%).

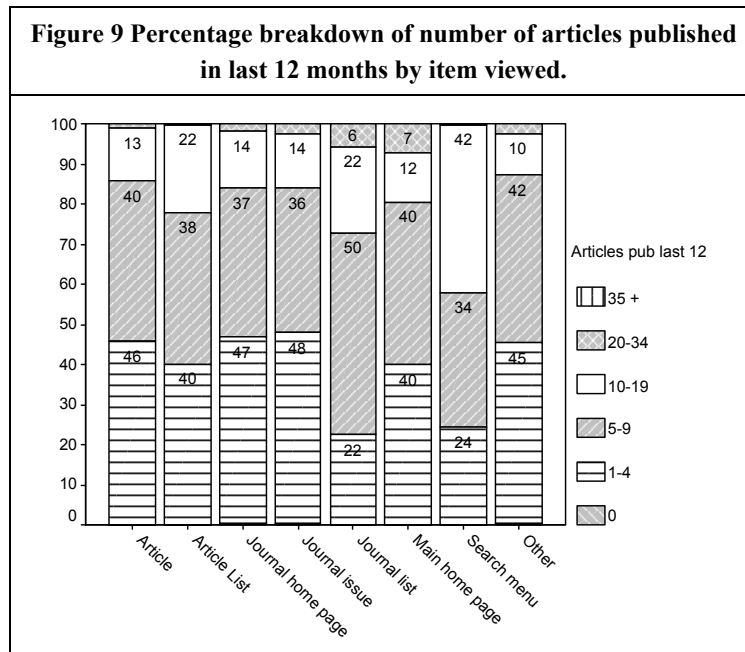


4.1.1.7 By number of articles published

In terms of the total numbers of articles respondents published, those using the Journal list and the main home page appear to have published the most (Figure 8). Maybe, these kinds of people were generally active and browse around a lot for data, ideas and opportunities?

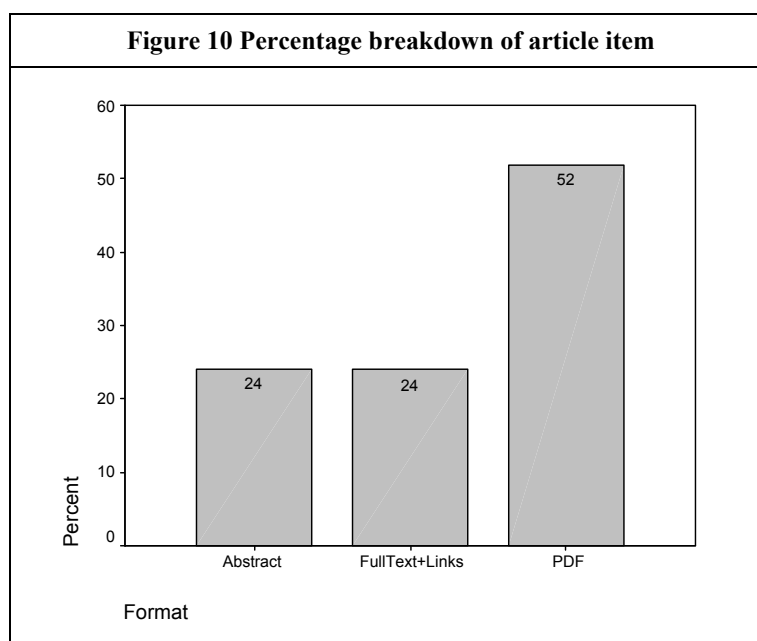


In terms of number of articles published in the last year those using the Journal list and the main search published the most (Figure 9).



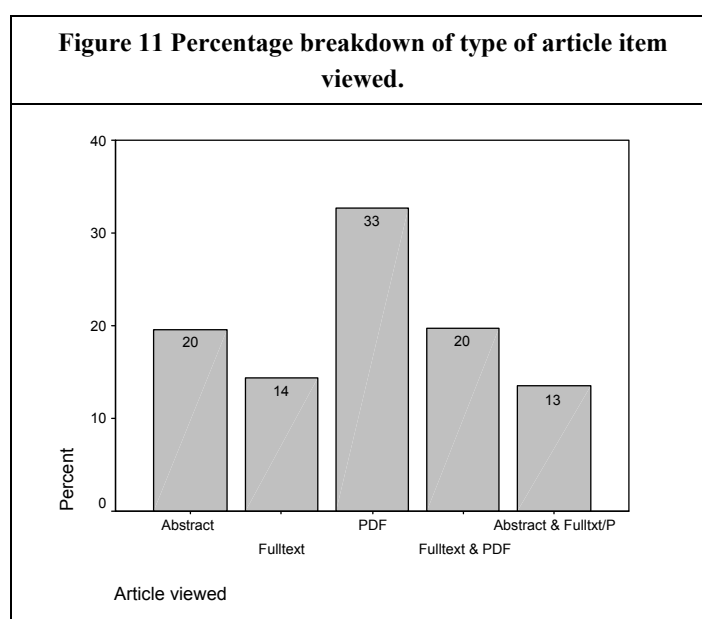
4.1.2 Type of article viewed

In the main the object of viewing ScienceDirect was to inspect or retrieve an abstract, full text or PDF or a combination of these three. In this section this is explored by examining views to Abstracts, the PDF version of the article and the HTML Full text version. The Full text (HTML) version also includes reference hotspot links that the user can click on to view additional articles. Links within HTML full text are citation and reference hotspot links that the user can click on to view related material. The figures for abstracts include summary plus (enhanced abstract) versions of the document this also includes links. About a quarter (24%) of views were made to abstracts a half (52%) to PDFs and a quarter (24%) to HTML Full text items, (Figure 10). The percentage of PDF views is twice as big as HTML full text views.



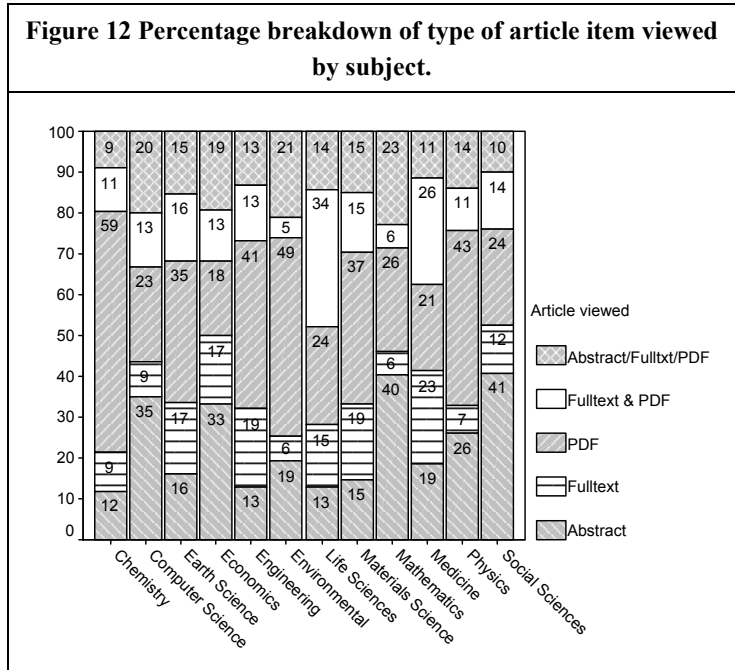
4.1.3 Session analysis

Previously we have been dealing with aggregated data; here we place this data in the context of a search session. Within a session users may view more than one item. Hence a distinction can be made between sessions where just an abstract was viewed and sessions that just viewed a PDF or a full text (HTML) item or a combination of them. Most (33%) sessions saw a PDF item being viewed; 20%, or one in 5 users, viewed a full text and a PDF item; the same proportion (20%) just viewed an abstract 14% just viewed a full text, and 13% viewed either an abstract and a full text or PDF item. Figure 11 relates.

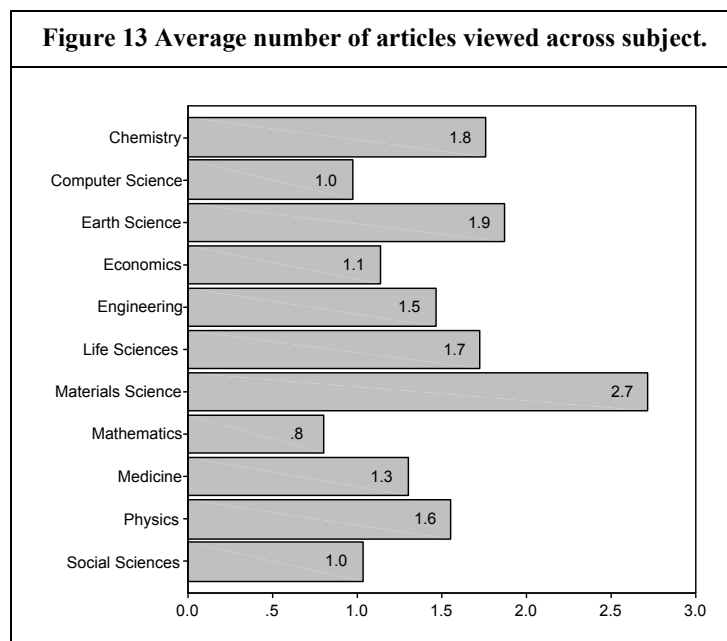


4.1.3.1 By subject back ground

In terms of subject groupings (Figure 12) Life Sciences recorded a higher than expected frequency of sessions just viewing a full text and PDF item (34%). Social Science respondents conducted the highest frequency of sessions just viewing abstracts items (41%); furthermore, Mathematics recorded above expected sessions only viewing abstracts (40%), as did Economics (33%) and Computer Science (35%). Chemistry recorded one the highest frequency of PDF item only sessions (59%); this was true, but to a lesser extent, of Environmental science (49%) and Engineering (41%). This merits further investigation.



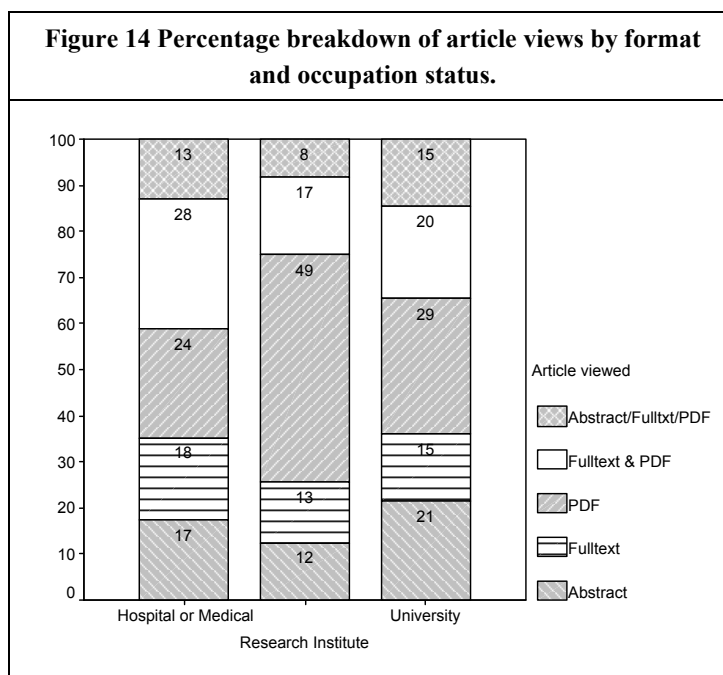
The average number of articles viewed in a session by the subject background varies quite considerably (Figure 13). Material Scientists recorded the highest average of 2.7 articles per session, while medical users recorded the lowest average of just 0.8 article views per session.



Authors as users: a deep log analysis

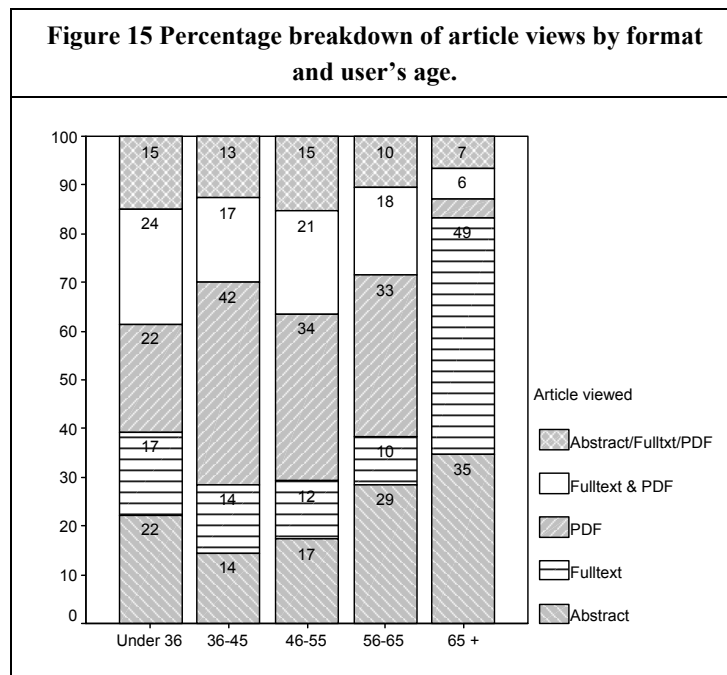
4.1.3.2 By type of organisation

Compared to universities, research institutes recorded greater use of PDF only sessions - 49% compared to 29% (Figure 14).



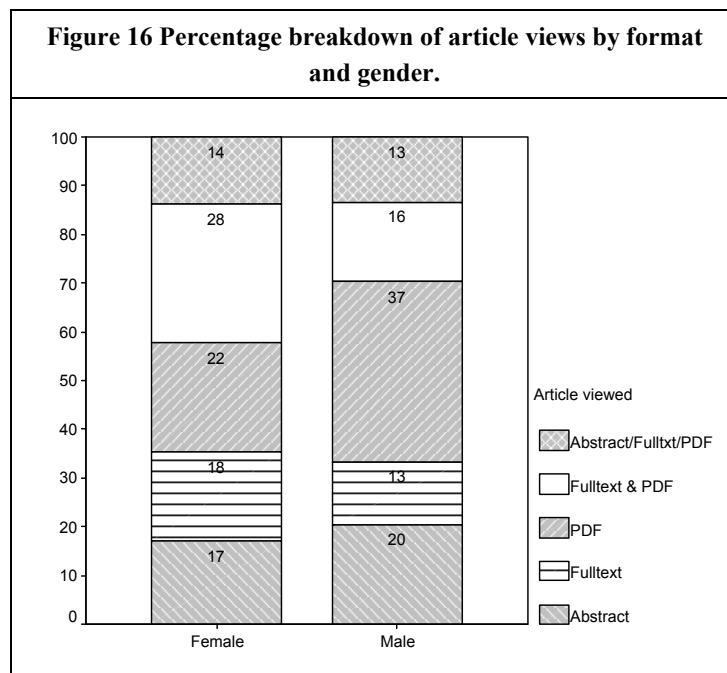
4.1.3.3 By age

The number of abstract only sessions, tended to increase with age, with the exception of the under 36s (Figure 15). About 14% of those 36 to 45 just undertook an abstract item session but this increased to 29% for those aged 56 to 65 and to 35% of those 65 and over. This could be a result of the time demands on senior people or it could be because with time they have learnt to determine relevance this way. The proportion of full text only sessions tended to decrease with age, they declined from 17% (under 36) to 10% (56 to 65).



4.1.3.4 *By gender*

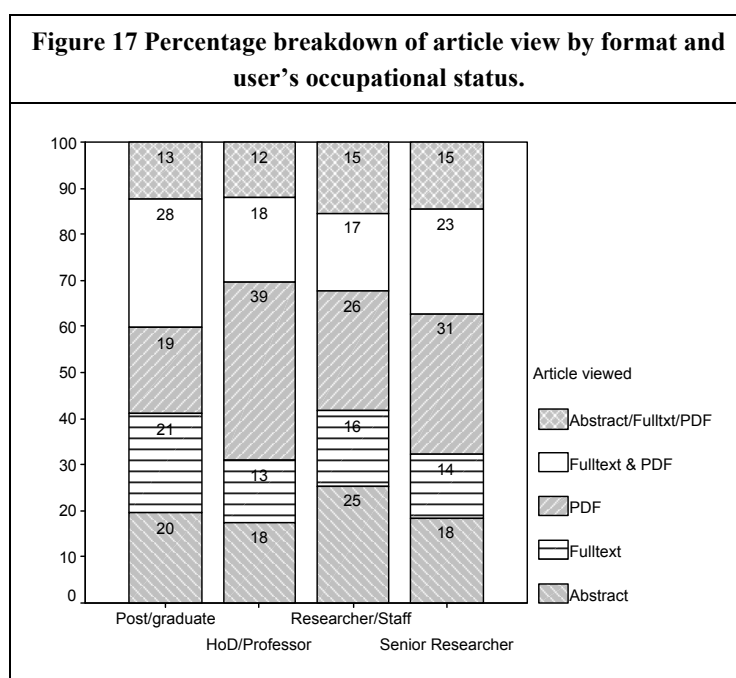
Men were much more likely to undertake a PDF only session -, 37% did so as compared to 22% for women. Women were more likely to have a Full text and PDF session, 28% compared to 16% for men (Figure 16). Gender has not been a significant variable generally but here it is and this merits further investigation.



Authors as users: a deep log analysis

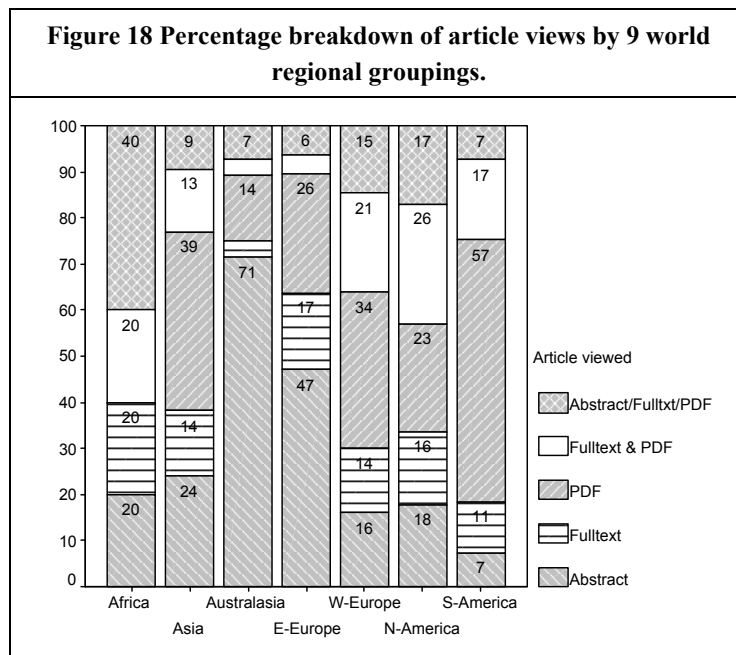
4.1.3.5 By occupational status

Students were more likely to record a fulltext only session (21%) or a full text or PDF session (28%) compared to other groups (Figure 17), while Heads of Department were more likely just to view PDFs, 39% of their sessions saw just PDFs viewed Senior researchers were more likely to view just abstracts (25%).



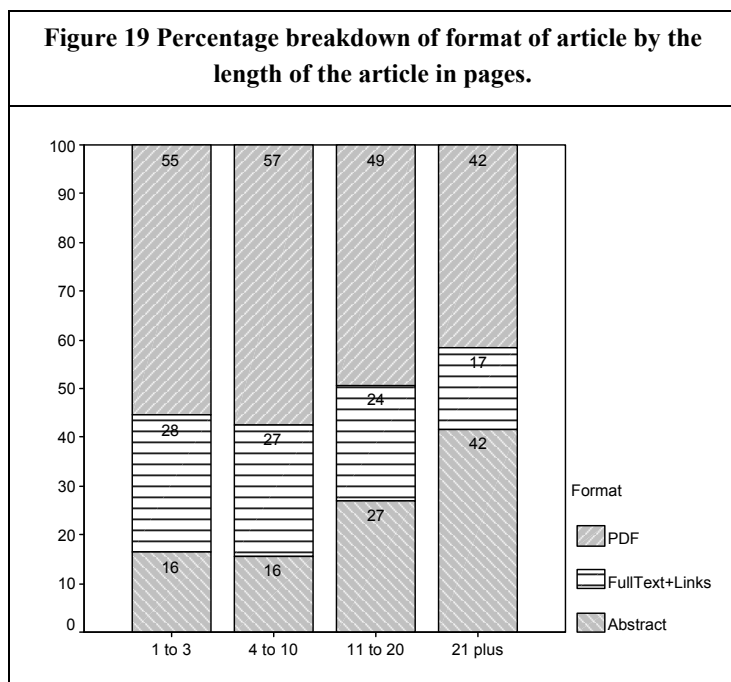
4.1.3.6 By geographical location

Users from Australasia were surprisingly heavy users of abstracts, approaching three quarters (71%) of their sessions viewed abstracts (Figure 18). Eastern Europeans were also proportionately big users of abstract only sessions (47%). South Americans recorded the highest proportion of PDF only sessions (57%).



4.1.4 Length of article viewed (number of pages in a paper)

Unusually, the number of pages for each article was also recorded in the logs. This gave us an opportunity to examine user behaviour in regard to length of article, long, short etc. The first analysis (Figure 19) examines the relationship between the number of pages in a paper and the format chosen to view the article. As we might have expected (but its good to have this confirmed), the longer the article, the greater the likelihood of viewing the article as an abstract or as a summary plus and the less likelihood of it being viewed in a PDF or Full text format. While 84%

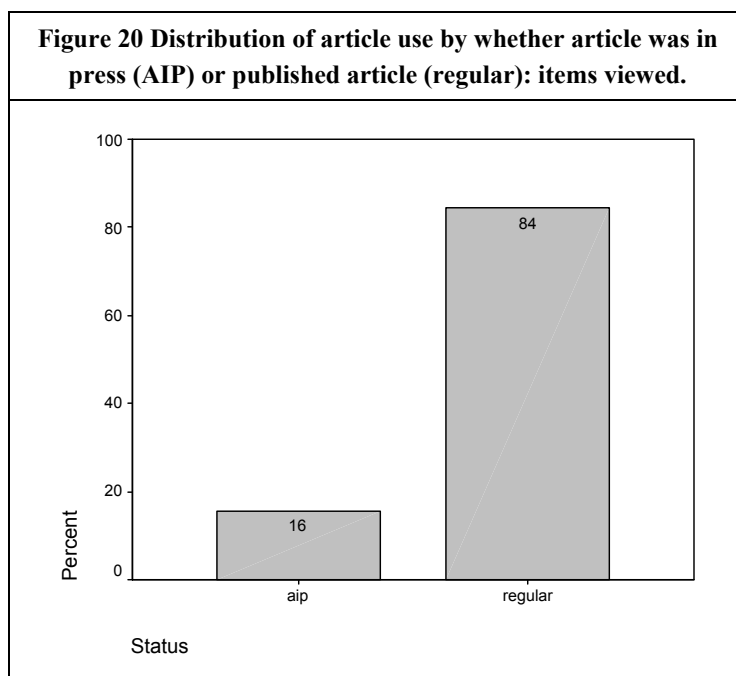


of papers less than 10 pages long were viewed as either a PDF (55%) or a Full text (28%), this falls to 73% for papers 11 to 20 pages long and to 58% of papers 21 or more pages long.

4.1.5 Publication status of article viewed

The ScienceDirect logs provided another opportunity to investigate a rather unusual deep log metric – the status of an article (pre-print or regular). Status refers to the print stage at which an article was at or its publishing status. Articles in press (AIP), refers to articles in press which have not been finalised but are available online. Regular describes finally published papers. The analysis tells us something about the need for currency on the part of the user.

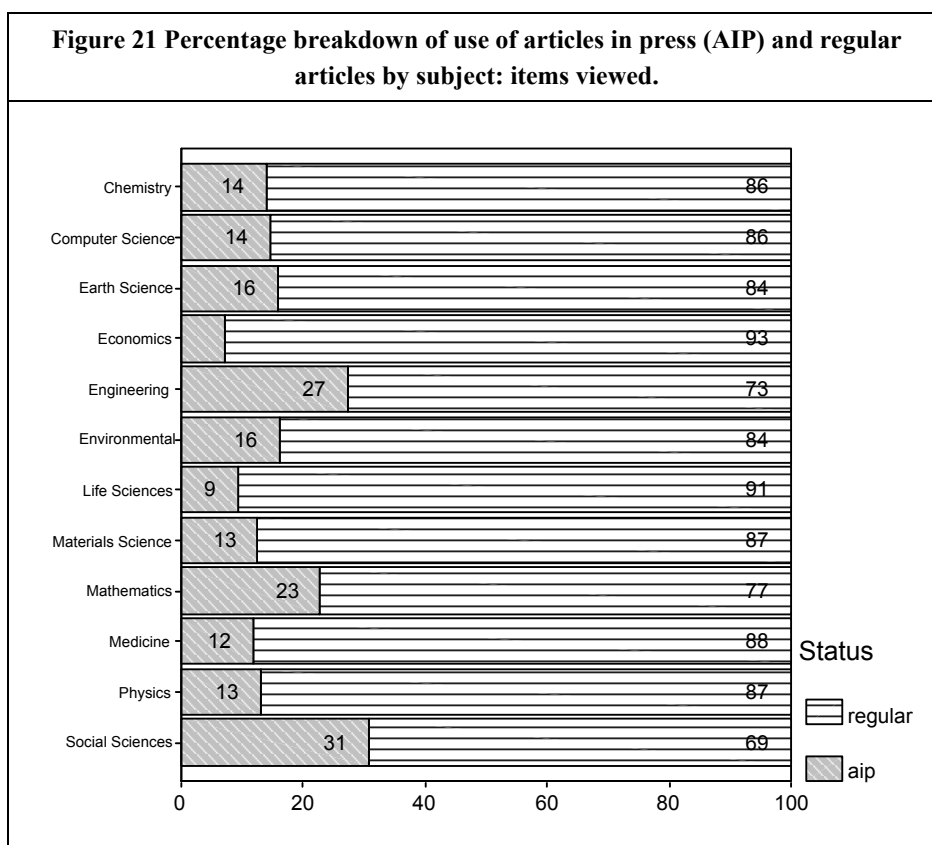
Sixteen percent of the total of article views were to articles in press and 84% to regular papers (Figure 20). In understanding the data and the relative popularity of AIPs it should be noted that nearly all journals have AIPs and that AIP is simply a stage in production and these versions get overwritten as vol/issue and page numbers are added to the article. Thus there is no real way of knowing at any given time what proportion they represent across all the ScienceDirect journals. However, anecdotally, based on a quick check of several journals, a journal is likely to have 5-10% of 1 years worth of articles in AIP form. This makes sense as AIPs will normally be compiled into an issue all at the same time - 5-10% is likely to be an issues worth. Once an issue is compiled AIPs will slowly build up again until the next issue is ready. It would seem they are then relatively popular and the currency they deliver probably explains this.



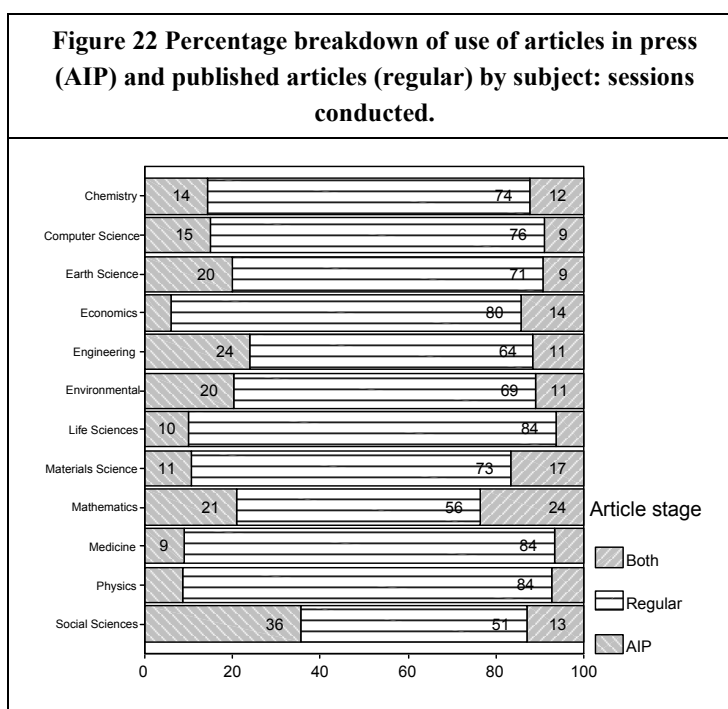
Most people viewed regular articles in their search session nearly three quarters (74%) of user sessions just viewed regular papers, 16% of session just viewed articles in press, while 11% of sessions viewed both articles in press and regular articles.

4.1.5.1 By subject

Users based in the Social Sciences (31%), Engineering (27%) and Mathematics (23%) recorded higher than expected views to articles in print (AIP) and this is clearly an indicator of the importance of currency in these fields (Figure 21). Economics (7%), Medicine (12%) and Life Science (9%) showed lower than expected use and this might be surprising, especially in the cases of medicine. An explanation might lie in the subject distribution of AIPs across the database.

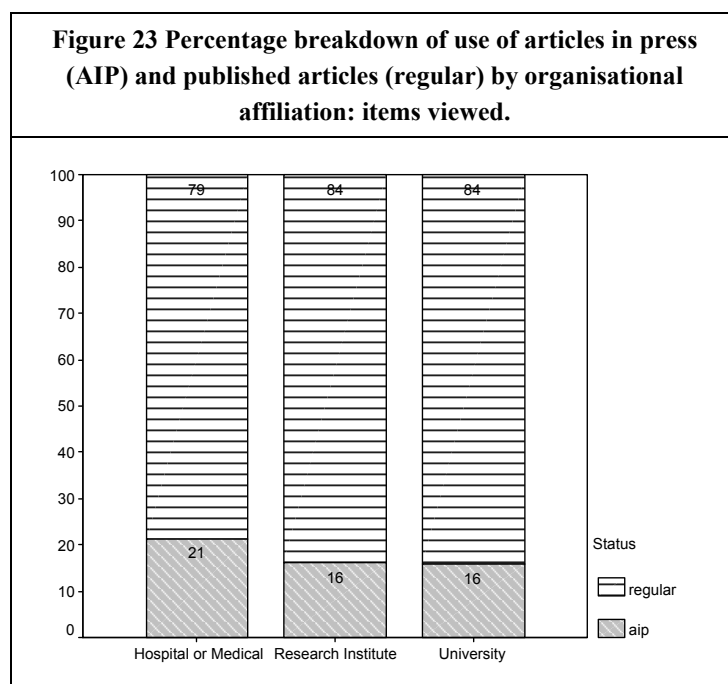


The Social Sciences (36%) and Engineering (24%) recorded the highest frequency of sessions that viewed only AIPs (Figure 22). Medical (84%) user sessions largely resulted in only views to finished articles (Regular), something which indicates the primacy of the finished article in the field.

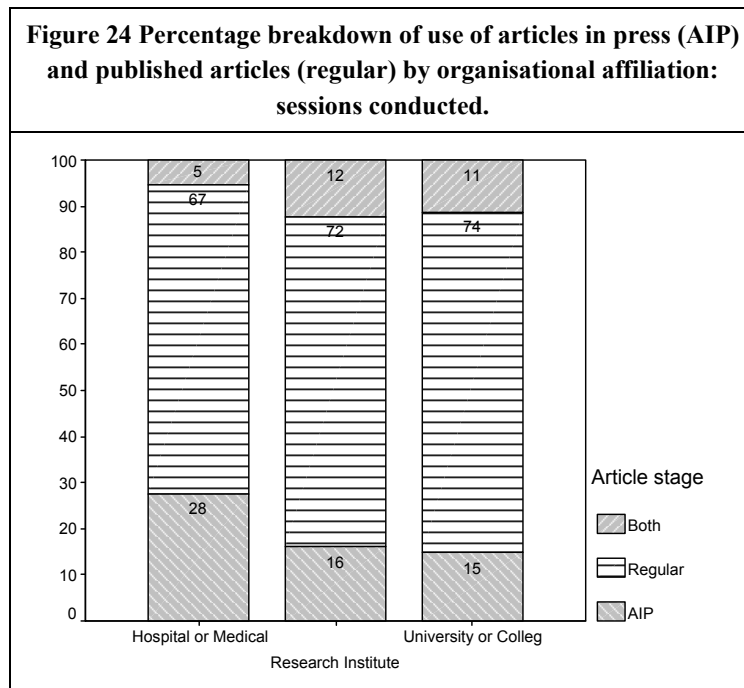


4.1.5.2 By type of organisation

Those based in hospitals were the biggest users of articles in print - 21% of their article views were to AIPs (Figure 23). Again, this is a possibly surprising result that requires further investigation. It contradicts the result above but may say something about the impact of location on use or the particular information seeking behaviour of practitioners.

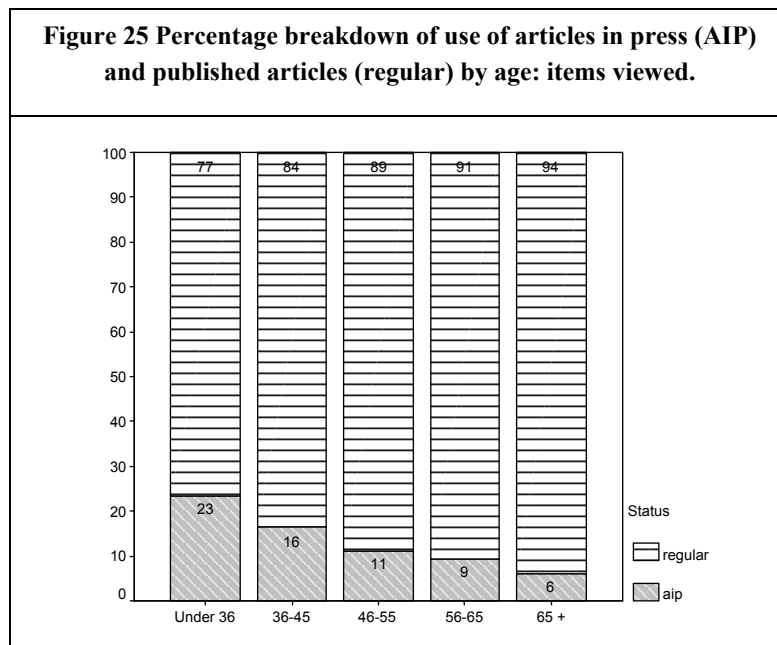


Users based in hospitals recorded the high number of AIP only sessions, 28% did so as compared to 15% for academics (Figure 24).



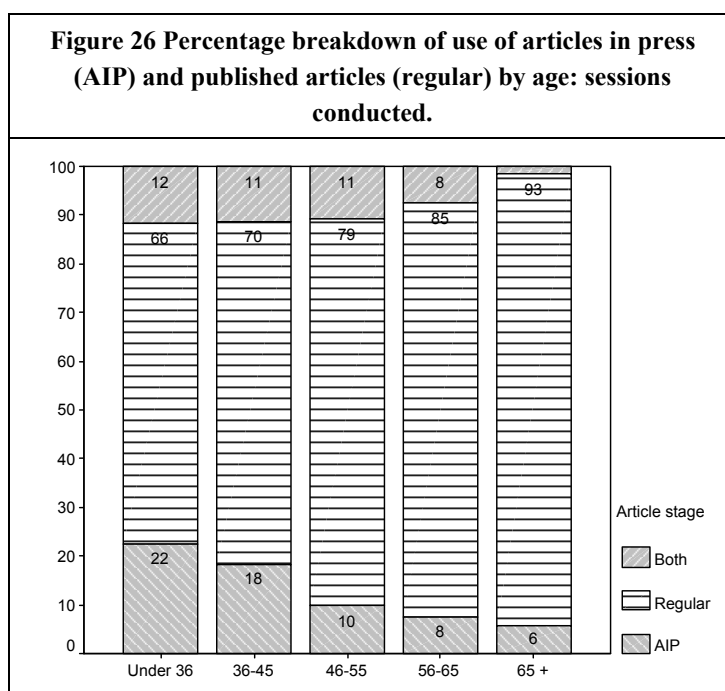
4.1.5.3 By age

An important relationship between age and use of articles in print was found. The younger the user, the more likely they were to use AIPs (Figure 25). About a quarter (23%) of those aged under 36 viewed articles in press articles, however this was only true of 6% for those aged over 65; quite a significant difference for which an explanation is not obvious.



Authors as users: a deep log analysis

The conduct of finished (regular) article only sessions tended to increase with age (Figure 26). The greatest frequency of AIP only sessions (23%) and joint AIP and (11%) sessions was in the under 35 age group. These percentages, in particular for AIP only sessions decreased with age with those over 65 being very unlikely to view articles in print. It is thought that older (and secure in tenure) researchers were more interested in saving time and hence may choose to disregard AIPs that are accessed by navigating hierarchical menu trees.

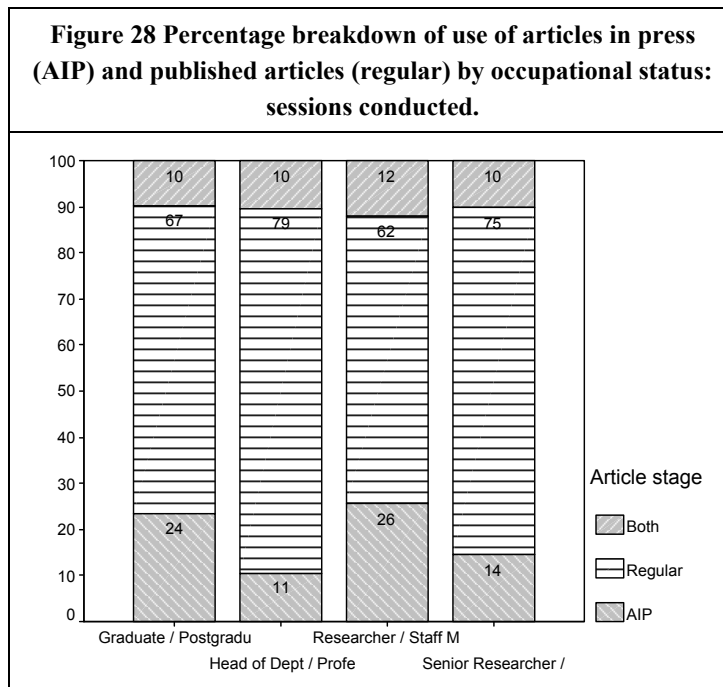
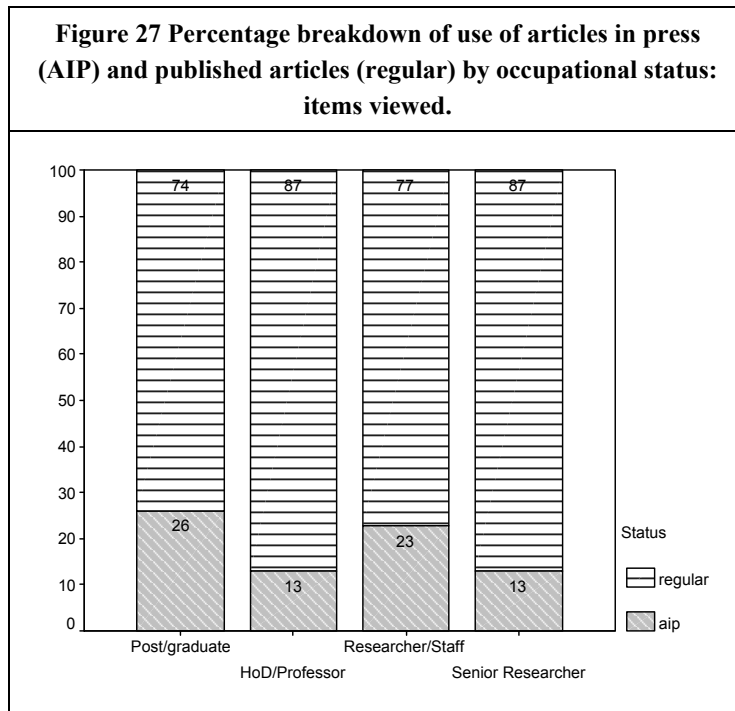


4.1.5.4 By gender

Men tended to make more views to articles in press than women, 18% compared to 10% and also were far more likely to have recorded an AIP only session; 19% did as compared to 6% for women.

4.1.5.5 By occupational status

In terms of occupational status, students made most use of AIPs – recording about a quarter (26%) of views, compared to 13% for Heads of Department, 13% for senior researchers and 23% for researcher/staff (Figure 27). This fits well with the age data (students are likely to be under 26). Again, this is another interesting finding which requires following-up. Researcher/staff members were most likely to construct a session were just an AIP (26%) or both (10%) a regular or an AIP paper were viewed. Students displayed a similar pattern (Figure 28). Both these groups recorded a higher than expected use of AIPs.

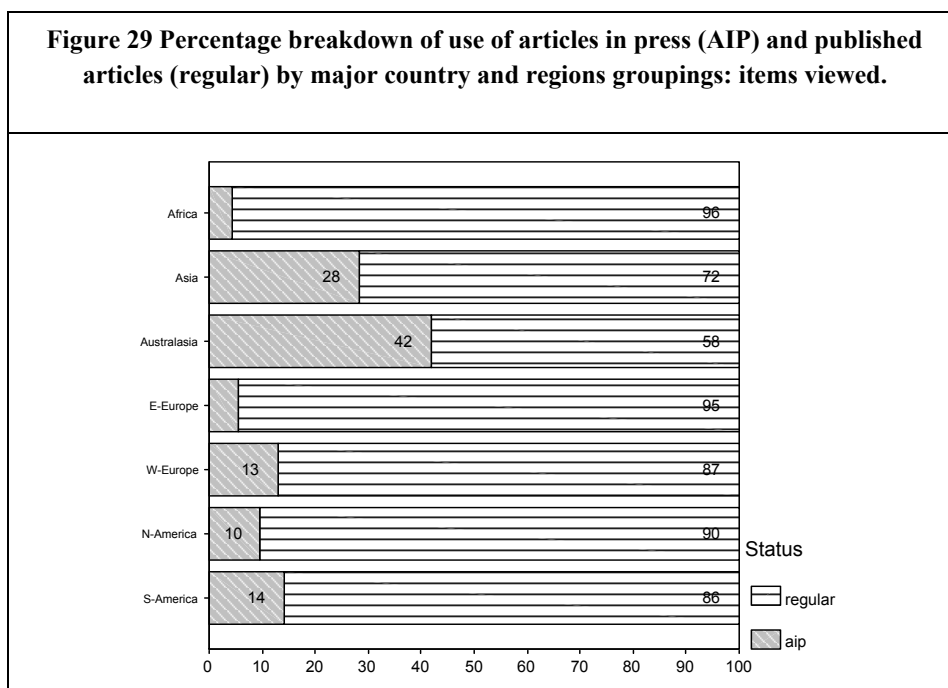


4.1.5.6 *By geographical location*

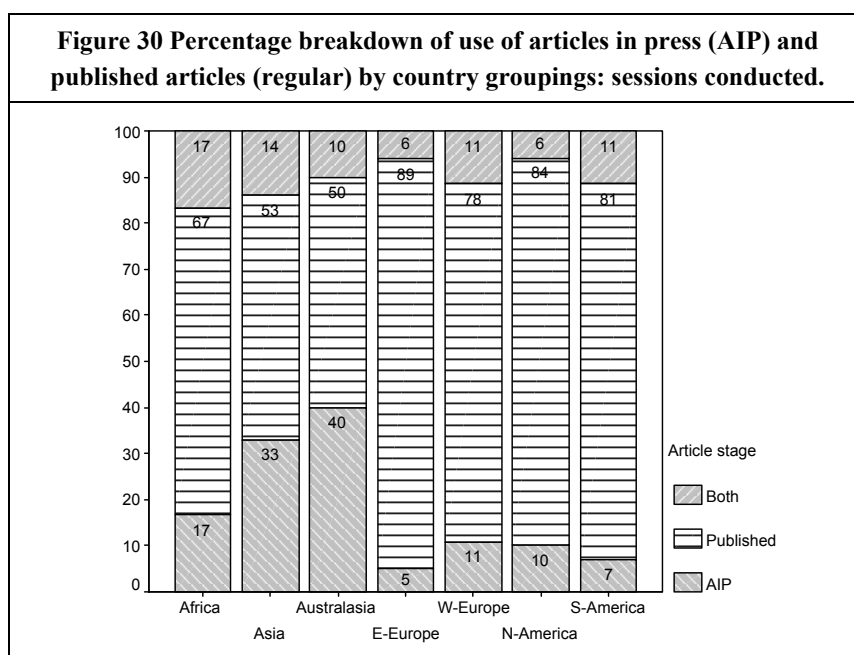
There were huge differences here (Figure 29); with users from Australasia recording the highest rate of views to articles in press (42%) and they were followed by Asian respondents

Authors as users: a deep log analysis

(28%). These are significant cultural differences for which no ready explanation comes to mind.

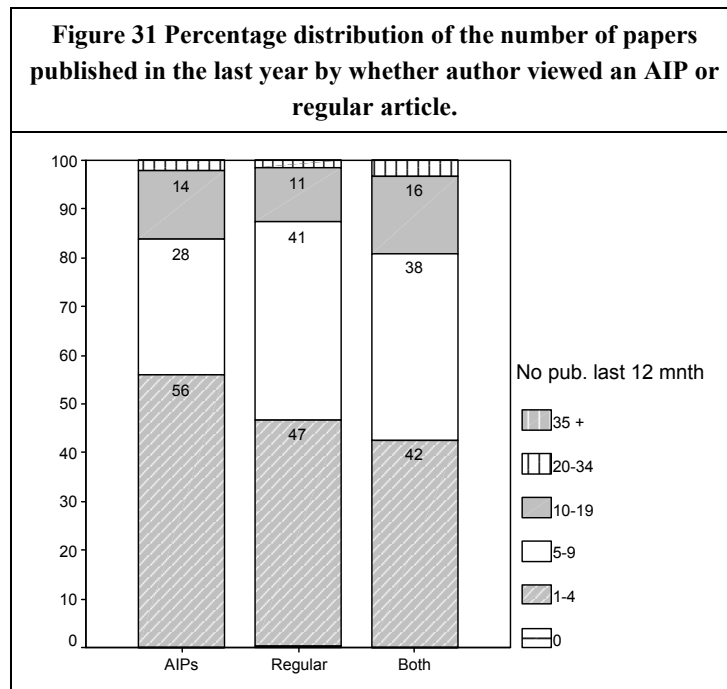


With regard to sessions (Figure 30) Australasian and Asian users, also undertook a higher than expected frequency of sessions where just articles in print were viewed (respectively, 40%, and 33%). Users from the North America and Eastern Europe also recorded the highest proportion of sessions where just regularly regular papers were viewed (84% and 89%).



4.1.5.7 *By number of articles published*

In terms of the number of papers published in the last year (Figure 31) those people just undertaking articles in-print sessions were far less likely to have published 5 or more papers - about 44% had done so as compared to 53% of those just viewing regular material and 58% of those users viewing both in a session.



4.1.6 **Publication year of article viewed**

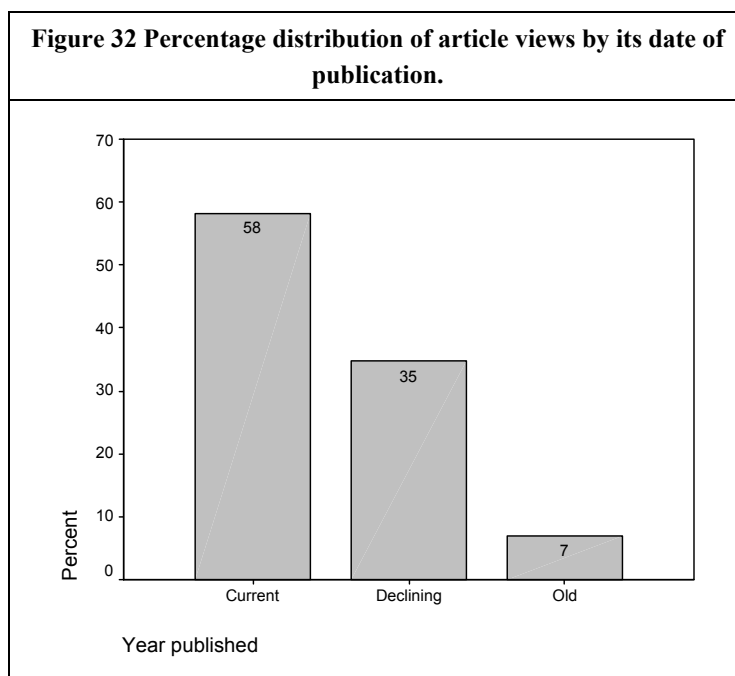
Generally for most journals articles were available from the early 1990s providing an archive of over ten years for many titles. However, some journals, like The Lancet, have an older back catalogue, in its case back to 1834. The year that the article was published was recorded and grouped into three time periods:

- Current – 2004 and 2005
- Declining – 1999 to 2003
- Old – 1993 to 1998

Well over half (58%) of the article items viewed were to the current one-year period (Figure 32), a third (35%) were to the period described as declining and only 7% were to articles 7 years or older. In terms of sessions most (55%) sessions just saw a current paper, 25% just viewed a declining aged paper, while 8% viewed at least one current and one

Authors as users: a deep log analysis

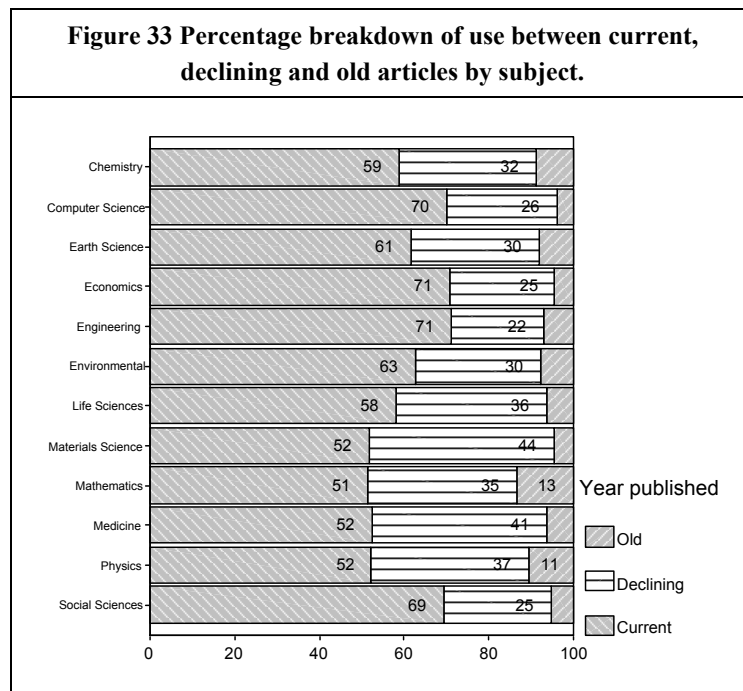
declining, and 12% viewed an old paper and or some combination of current and declining in their paper.



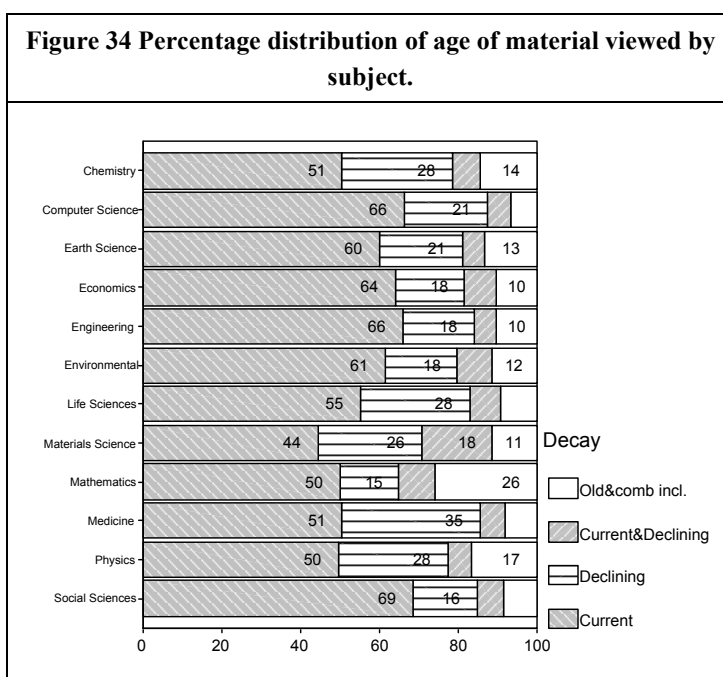
4.1.6.1 By subject background

Users from Economics (71%), Engineering (71%), the Social Sciences (69%) and Computer Science (70%) made above expected views to items to the current period (Figure 33). Respondents in these fields have previously been noted as recording high views to AIP papers and it appears that respondents associated with these subjects were particularly interested in current material. In terms of the use of older material, Mathematics (13%) and Physics (11%) scored highly. Respondents from Material Science (44%), Medicine (41) recorded relatively high usage of declining material.

Our previous work on journal databases has shown that search engines has provided improved access to older material and this has resulted in a greater than expected use of older material. This does not appear to be the case with ScienceDirect and this could be due to the higher scientific content of the database as compared to Synergy and OhioLINK, or it could be to do with the way people search the database. What appears to have happened is that there is more of a subject level playing field with areas traditionally associated with the use of older material (like the arts and humanities) showing higher levels of current use than fields associated with current use, like physics. However, with so few respondents for the arts and humanities (14) we should not jump to firm conclusions yet.

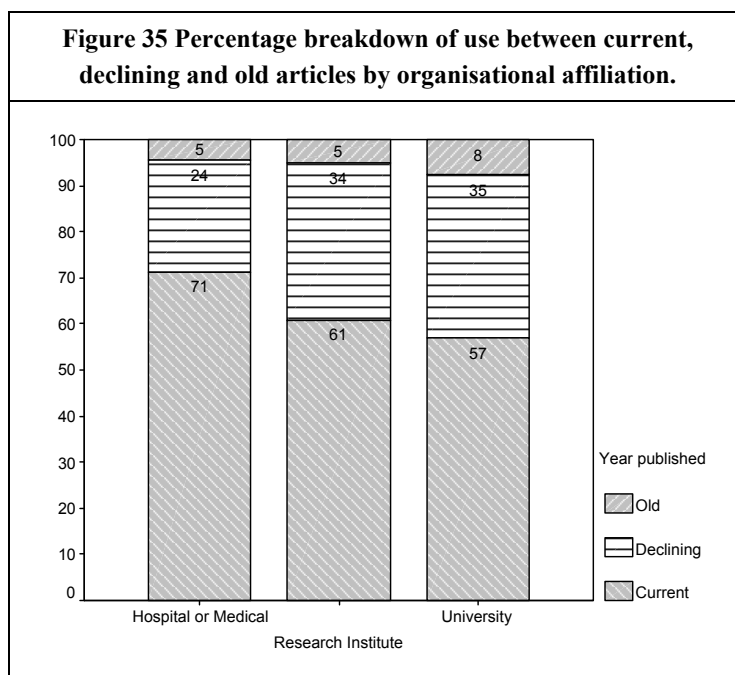


In terms of sessions conducted that included a view to old material (Figure 34), Mathematics (26%) and Chemistry (14%) and Physics (17%) recorded relatively high figures (Figure 34). And users from these subjects tended to have fewer sessions where just current material was viewed (respectively, 50%, 51% and 50%). Social Scientists were more interested in current material and about two thirds (69%) undertook sessions where just current material was viewed. Social Science users as we have already seen were particularly interested in articles in print, hence currency seems to be important. This appears to be at odds with expectation and could mark changes in behaviour as a result of the increased accessibility to older items. Users from Engineering (66%) and Computer Science (66%) also tended to record sessions where just current material was viewed.

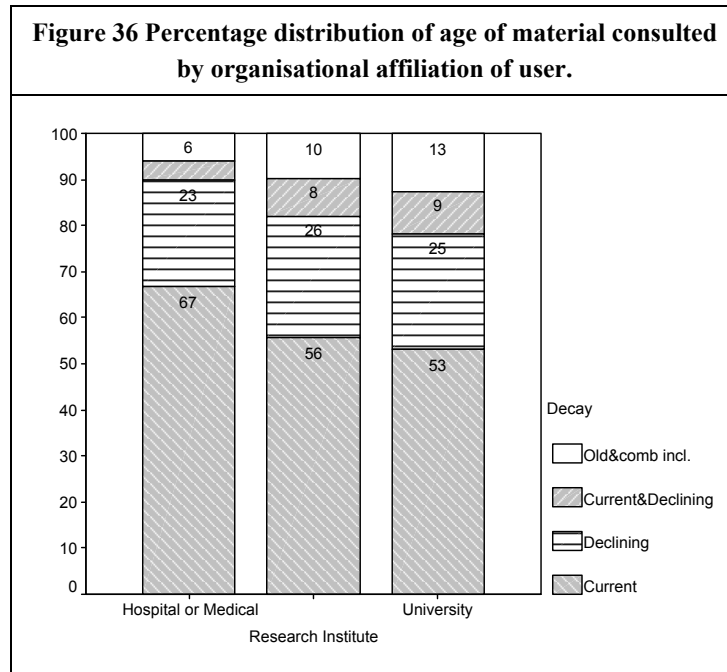


4.1.6.2 By type of organisation

Respondents based in Hospitals (Figure 35) were far more likely to view current articles (71%) while those based in Universities made good use of declining and old material (43%). In terms of sessions undertaken two-thirds (67%) of the sessions of hospital staff just saw current material being viewed, while users from universities and colleges (47%) and research institutes (44%) recorded sessions that included either declining or old material (Figure 36).

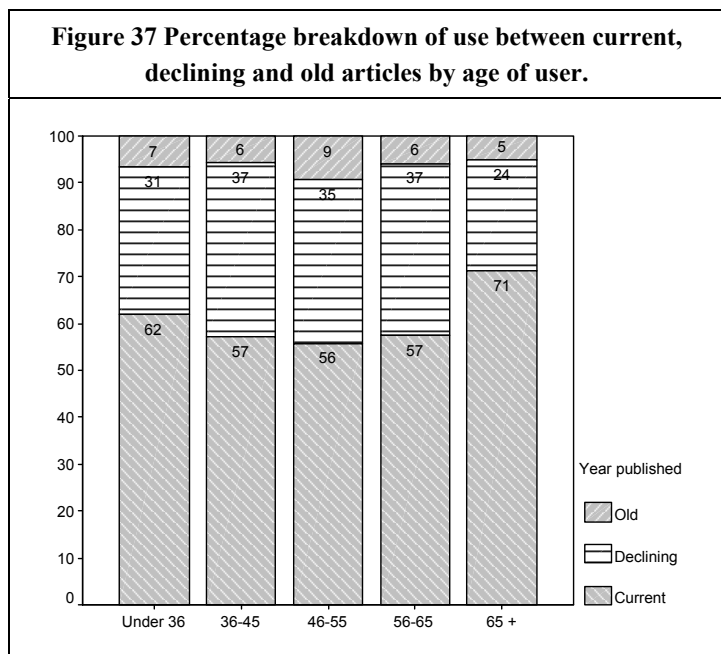


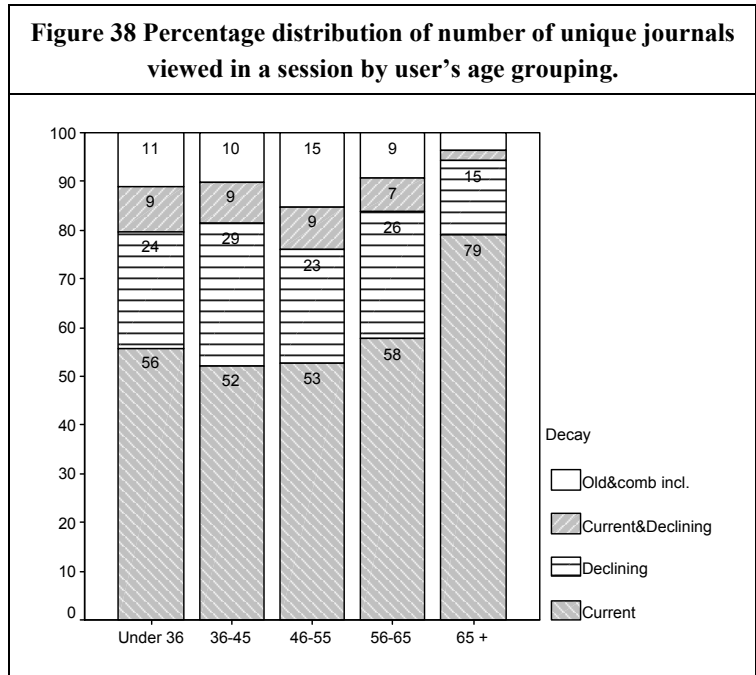
In terms of sessions undertaken two-thirds (67%) of the sessions of hospital staff just saw current material being viewed, while users from universities and colleges (47%) and research institutes (44%) recorded sessions that included either declining or old material (Figure 36).



4.1.6.3 By age

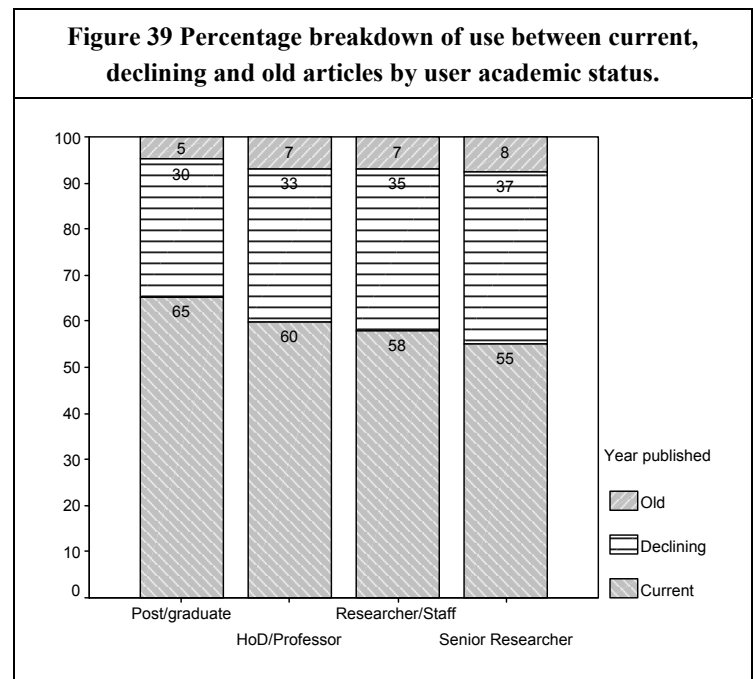
There is a marked curve in the use of current material by age (Figure 37), with older and younger users more likely to look at such material while those aged 36 to 55 were proportionately more likely to view older material; about 44% of items viewed by this age group were declining or old. Nearly half (48%) of those aged 36 to 55 conducted search sessions where declining or old material was viewed (Figure 38). Those aged 46 to 55 appeared most interested in old material and 15% undertook sessions that included views to older material. Young (under 36) and older (over 65) users were more like to conduct sessions where just current material was viewed.



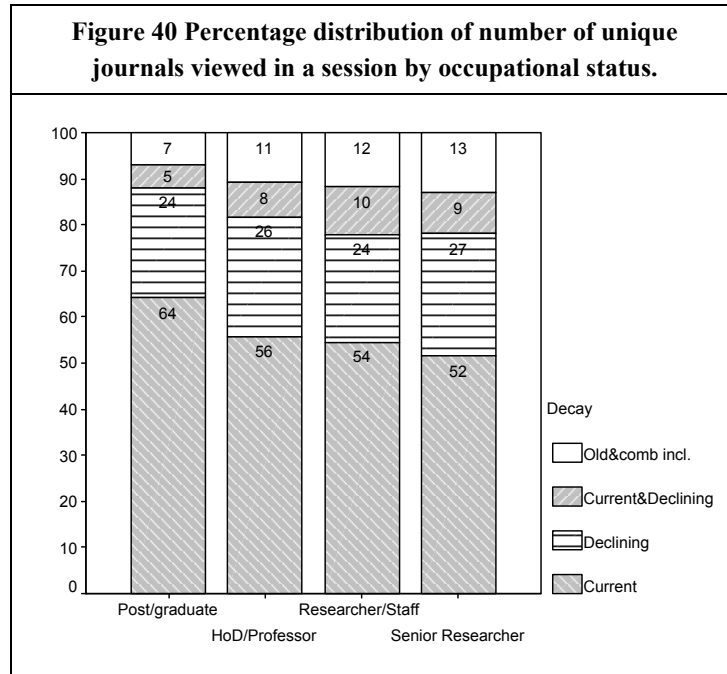


4.1.6.4 *By occupational status*

Senior researchers made more views to historical material (Figure 39), with 37% of their views featuring declining material and 8% old material.

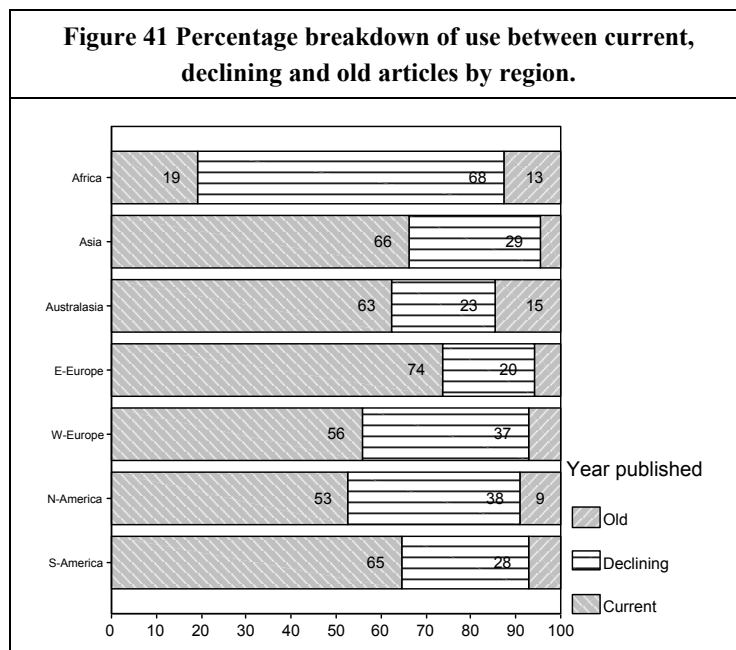


Regarding search sessions (Figure 40), postgraduate students were most likely to just view current material in a session, 65% did so and senior researchers least likely to (52%).



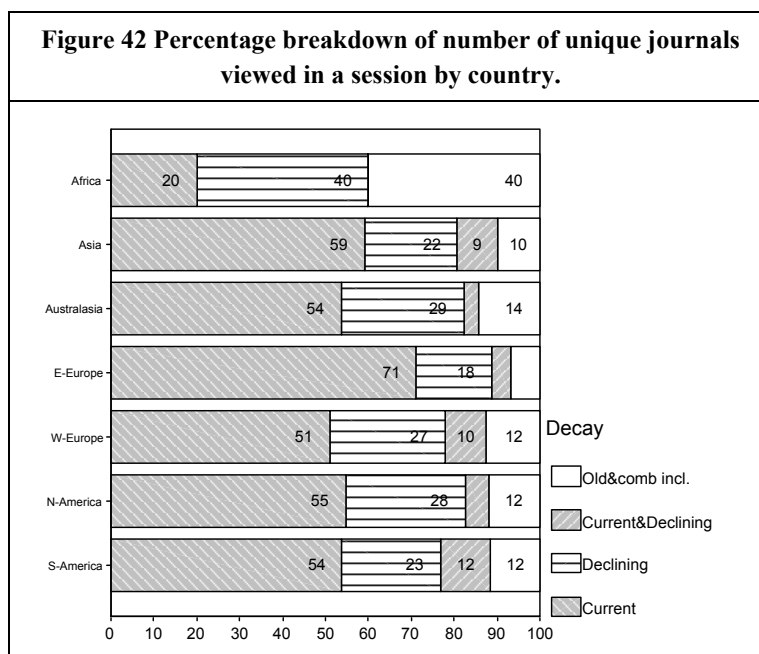
4.1.6.5 By geographical location

Of the regions, Eastern European (74%) and South American users (65%) recorded the highest percentage of views to current material (Figure 41). Australasian (15%) and North American users (11%) recorded the highest relative use of old material.



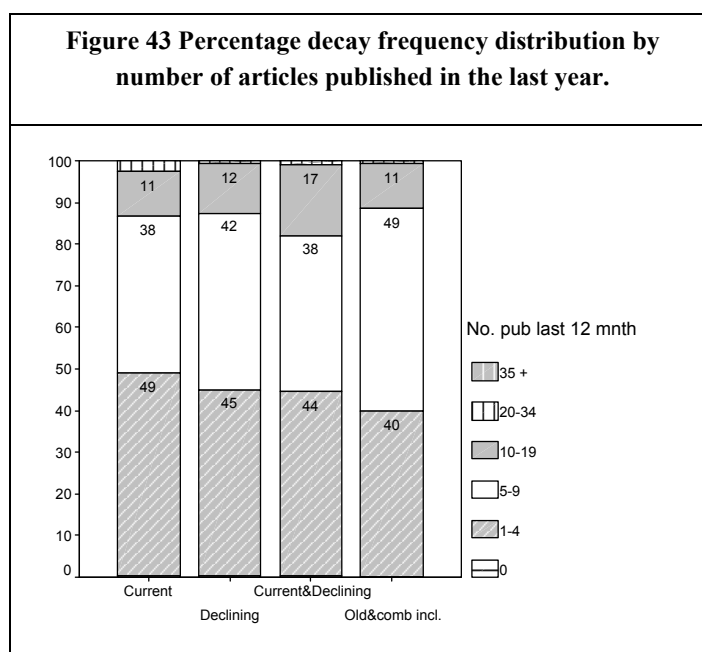
Authors as users: a deep log analysis

Over half of the users from North America (56%) viewed current material only in a session, 28% just viewed declining material and 17% had a session where current, declining or old material was viewed (Figure 42). Users from Eastern Europe (71%) and Asia (60%) conducted high levels of sessions just viewing current material.



4.1.6.6 By number of articles published

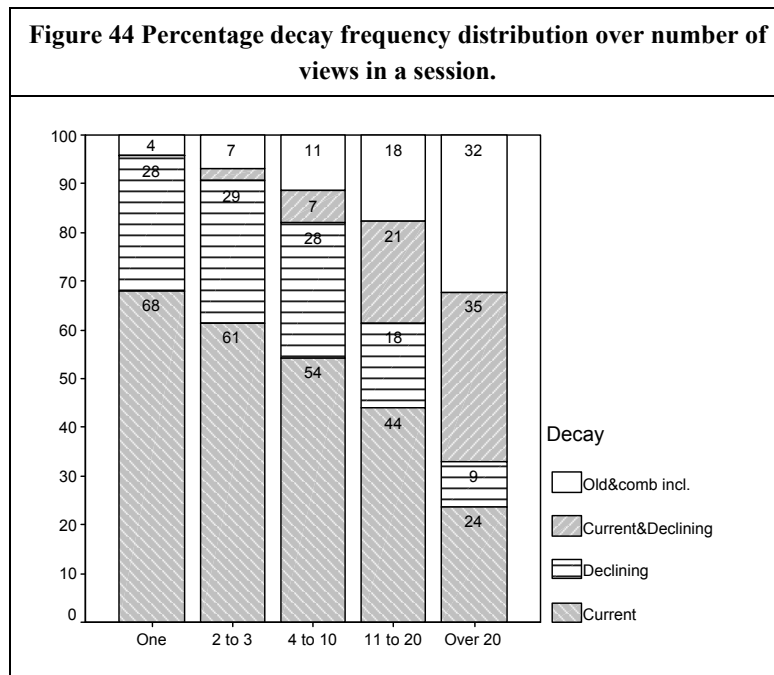
In terms of number of papers published in the last year those people viewing a wider range of historical material tended to publish more: 60% of those viewing a combination of



aged material, including old, had published 5 or more papers and this compares to about 50% who had done so for those users who had just viewed current material in a session (Figure 43).

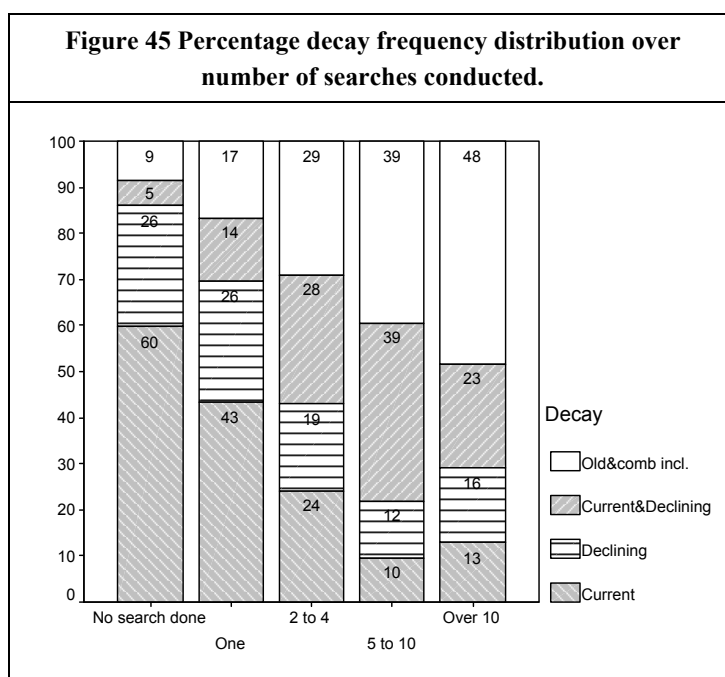
4.1.6.7 *By number of views made in a session (Site penetration)*

In terms of number of views in a session (Figure 44) we would expect those executing longer sessions to view a greater range of historical material and indeed this is the case. About a two thirds (61 to 68%) of views by those just viewing 3 or less items just looked at current material; however, this was true of just 24% of those viewing over 20 items.



4.1.6.8 *By number of searches undertaken*

Those users conducting more searches were more likely to view a wider date range of material (Figure 45). Clearly we would expect this, as more searches would argue for a greater historic range in article item viewing.



4.1.7 Individual journal titles used

Table 2 gives the top 20 journals by the number of items viewed and in general usage was widely distributed at the top, with the exception of the European Journal of Operational Research which recorded 4% of article item views (including journal menu views) and this surely needs an explanation as the performance is puzzling. With 20 of the 1700 or so journals available on ScienceDirect attracting nearly 21% of use there is the inevitable concentration in use that is so characteristic of journal ranked lists

Table 2: Top 20 journals (all article items and including journal menu views).

Journal	Number	%
European Journal of Operational Research	3041	4.0
Nuclear Instruments and Methods in Physics	992	1.3
Phytochemistry	926	1.2
Tetrahedron Letters	920	1.2
Journal of Alloys and Compounds	765	1.0
Powder Technology	760	1.0
Journal of Solid State Chemistry	712	.9
Earth and Planetary Science Letters	697	.9
Mathematical and Computer Modelling	666	.9
Biomaterials	646	.8
Biochemical and Biophysical Research Com	627	.8

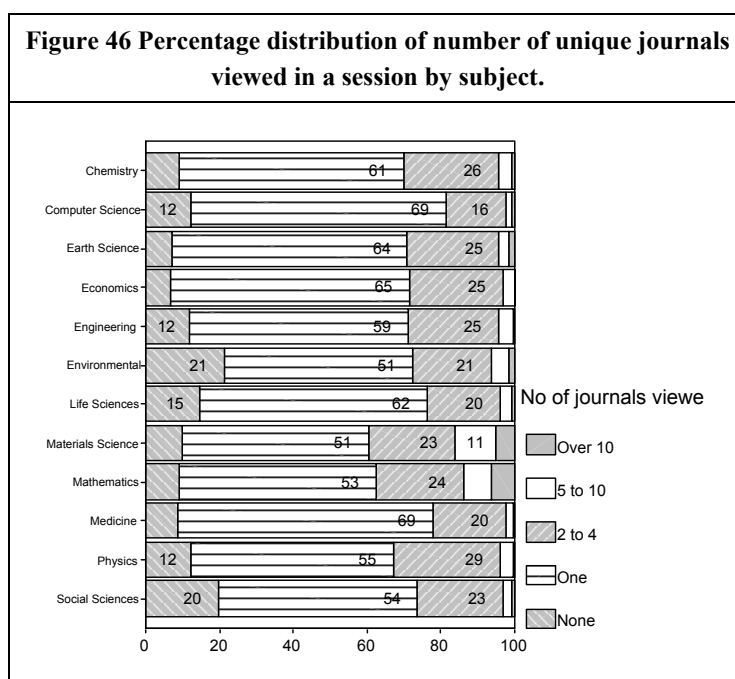
Physica C: Superconductivity	607	.8
Materials Science and Engineering A	580	.8
Journal of Catalysis	571	.7
Operations Research Letters	568	.7
Tetrahedron	560	.7
Thin Solid Films	551	.7
Journal of Nuclear Materials	539	.7
Applied Catalysis B: Environmental	525	.7
Journal of Molecular Spectroscopy	522	.7
As a % of total.		20.7%

4.1.8 Number of unique journals viewed in a session.

Most sessions (59%) saw just one journal being viewed, which is interesting given the sheer number of journals of offer, but maybe users know what they are looking for or just do not have the time to range widely. 23% viewed 2 to 4 journals, 13% didn't view any journals, 4% of sessions viewed between 5 to 10 different journals, while about 1% viewed over 10. Users can alternatively just view search results or menu items.

4.1.8.1 By subject background

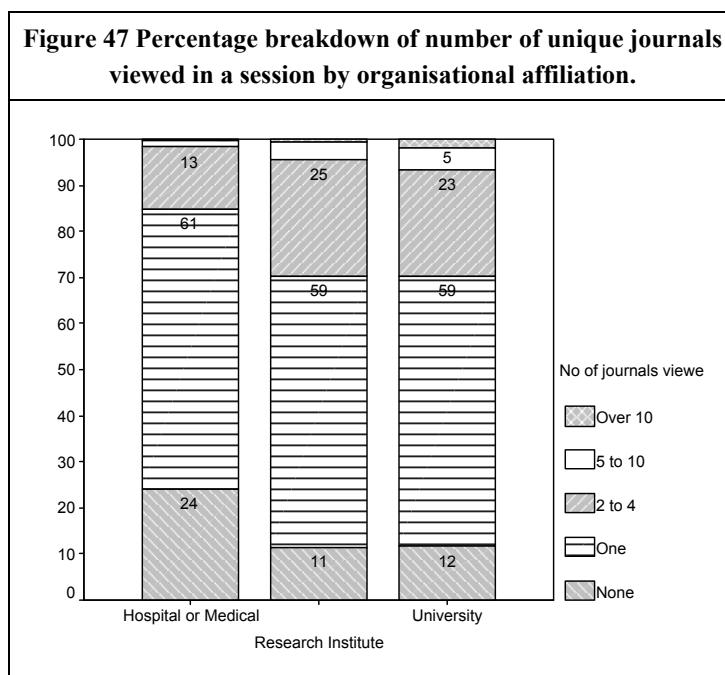
Users from Material science (39%), Mathematics (38%), Physics (33%) and Chemistry (32%) were most likely to view 2 or more journals in a session (Figure 46). While users from Medicine (69%) Computer science (69%), and Life Sciences (62%) tended to view just one journal in a session.



Authors as users: a deep log analysis

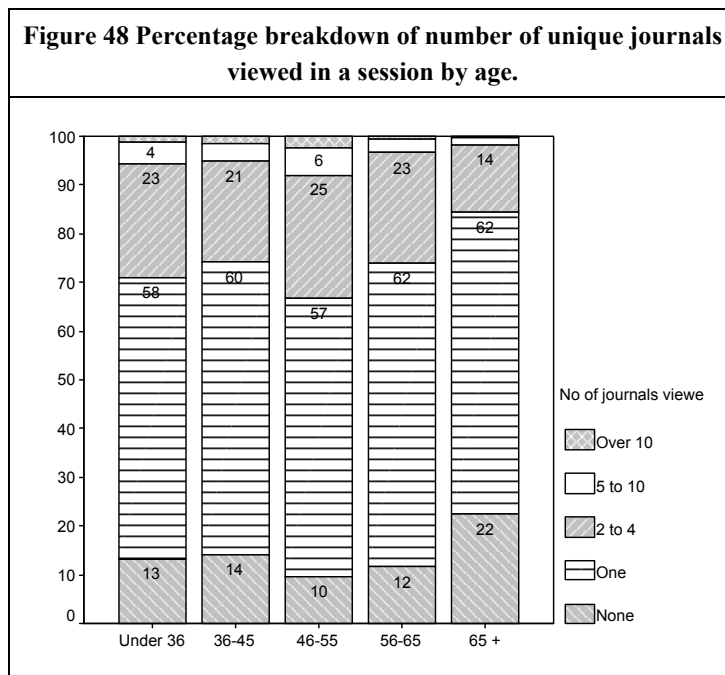
4.1.8.2 By type of organisation

In terms of where the user worked (Figure 47) those users based in hospitals were least likely to view a journal (15%) and were least likely to view 2 or more journals in a session; suggesting perhaps that the journal brand is less important here.



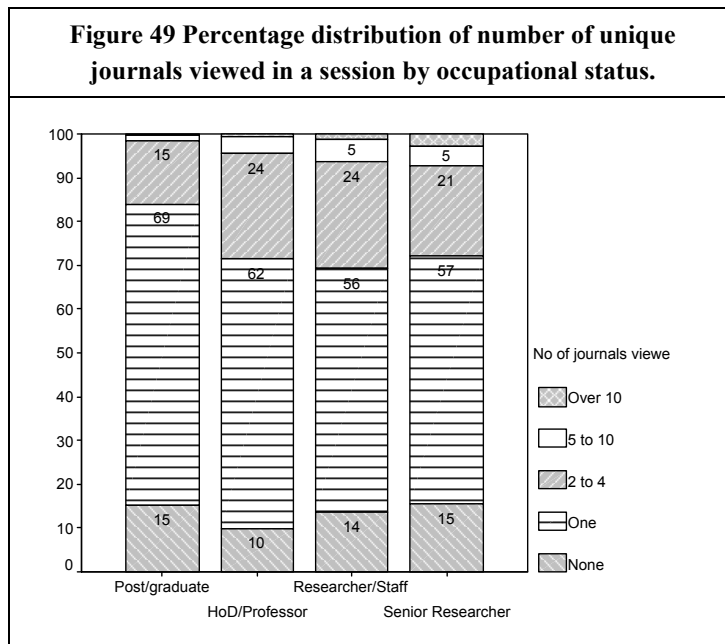
4.1.8.3 By age

Those aged 46 to 55 were most likely to view a journal - 90% did so (Figure 48), and were most likely to view 2 or more journals, a third did so: suggesting that this age group is the most active in reading material. Those aged over 65 were likely not to view a journal item (22%) and were least likely to view 2 or more journals in their session.



4.1.8.4 By occupational status

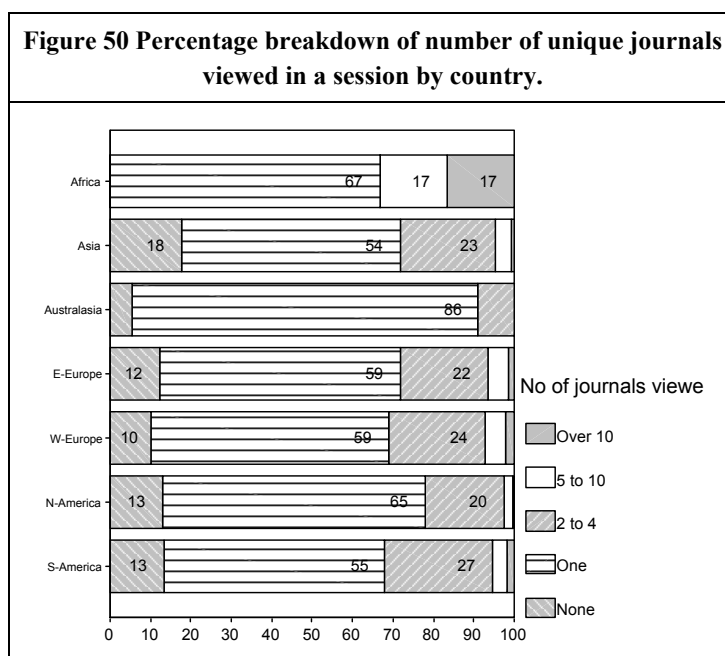
Students (Figure 49) were most likely to view just one journal (69%) while researchers (staff) were most likely to view 2 or more journals (30%). This supports our thesis of people checking up on references supplied to them. We have found in our other studies that those using the search facility are more likely to view more journals and it of interest to note that students seem less able to interface with the search facility compared to researchers.



Authors as users: a deep log analysis

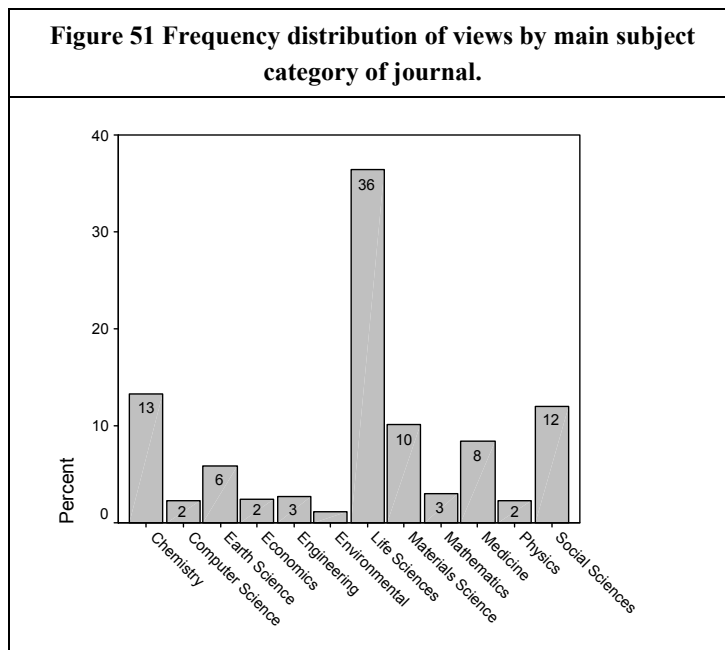
4.1.8.5 By geographical location

Australians (9%) North American respondents (22%) were least likely to view 2 or more journals in a session, while those from Western Europe (31%) were the most likely to (Figure 50).



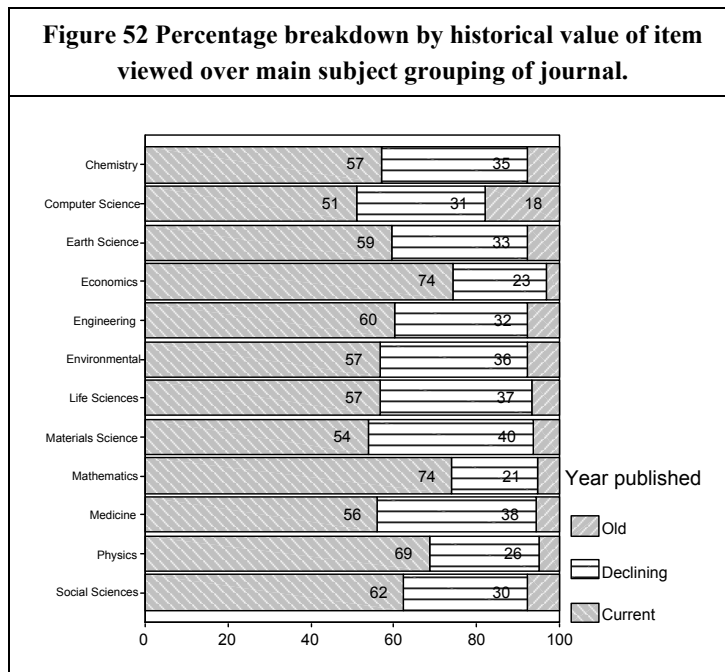
4.1.9 Subject of journals viewed

It was possible to categorise the subject of use by journal subject category. Figure 51 shows journal usage by main journal subject. Most use (36%) was attributed to Life Science journals; Material science accounted for 10% of usage and Chemistry 13%.



4.1.9.1 By age of article item viewed

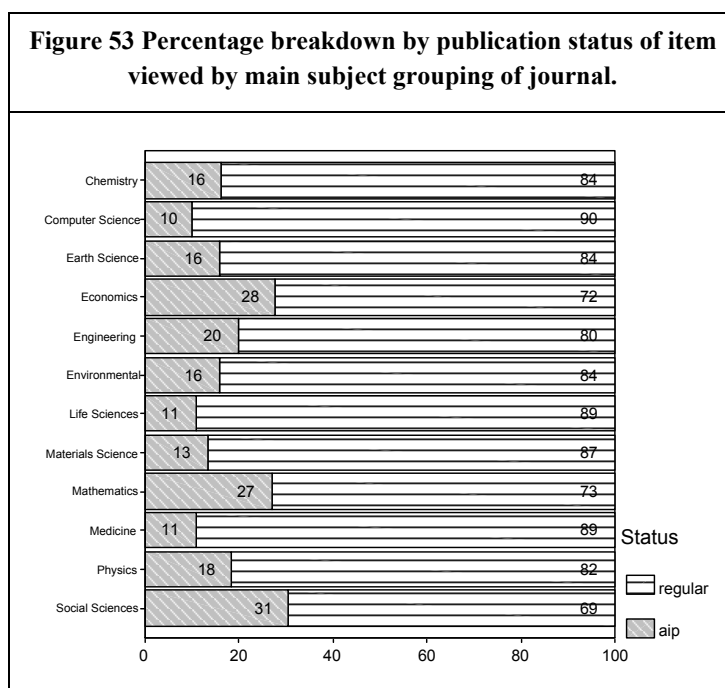
In terms of decay of article item viewed (Figure 52) subjects with the highest views to current material were Mathematics (74%), Social Sciences (62%) and Economics (74%), while Material Science (54%), and Computer Science (51%) recorded low views to current material. Oddly Computer Science journals scored the highest views (18%) to old material.



Authors as users: a deep log analysis

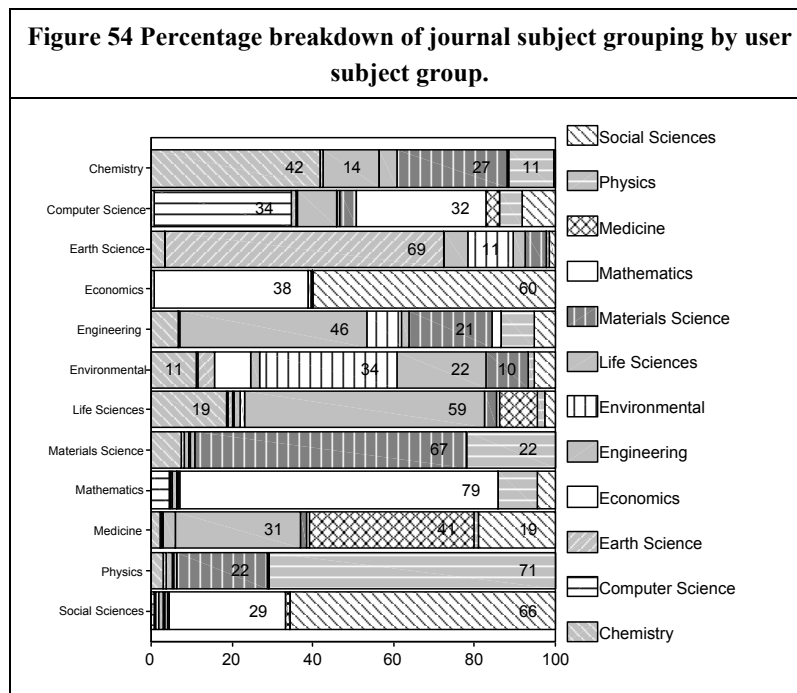
4.1.9.2 By publication status of article viewed

Figure 53 shows use of articles in press for the broad subject groups. In terms of views to AIP items Social Science (31%), Economics (28%) and Mathematics (27%) journals scored highly while Life Science (11%), Compute Science (10%) and Medicine (11%) scored lowly.



4.1.9.3 By subject background

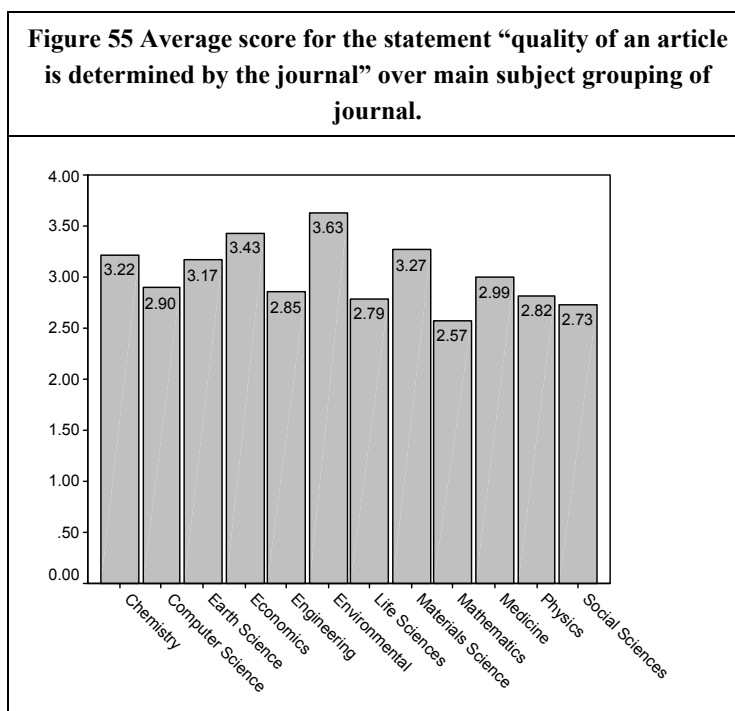
In this section we relate user subject background to the subject of journals used. In the main users of a discipline tended to use the journals related to that discipline (Figure 54).



Thus, 71% of those describing themselves as coming from physics viewed physics journals and this subject recorded the highest accord between user discipline and journal discipline. Environmental and Computer Sciences were least likely just to view journals within their discipline. Computer Science users were just as likely, 32% did, to view Mathematic journals while Environmental Science users (22%) also viewed Life Science Journals.

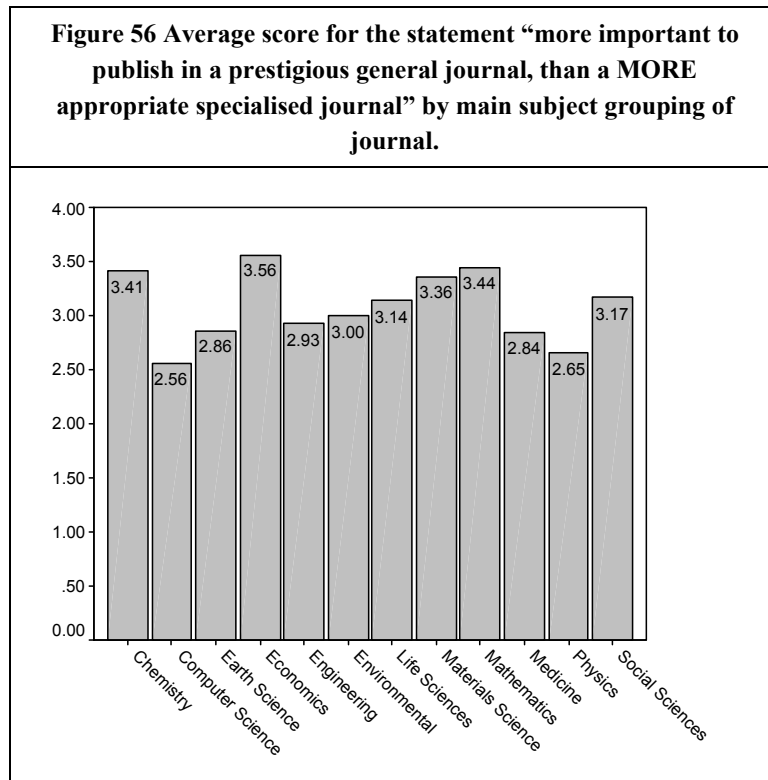
4.1.9.4 By whether respondents thought the quality of an article was determined by the journal in which it was published

We now relate attitudinal data with subject of use data. Figure 55 gives the average score for the question “the quality of an article is determined by the journal” by use of journal subject grouping. There is not a general consensus. Those respondents viewing Environmental Science (3.6) and Material Science (3.3) journals were more likely to agree with this statement, while Mathematics (2.6), Life Science (2.8), Engineering (2.9) and Computer Science (2.9) users appeared more likely to disagree.



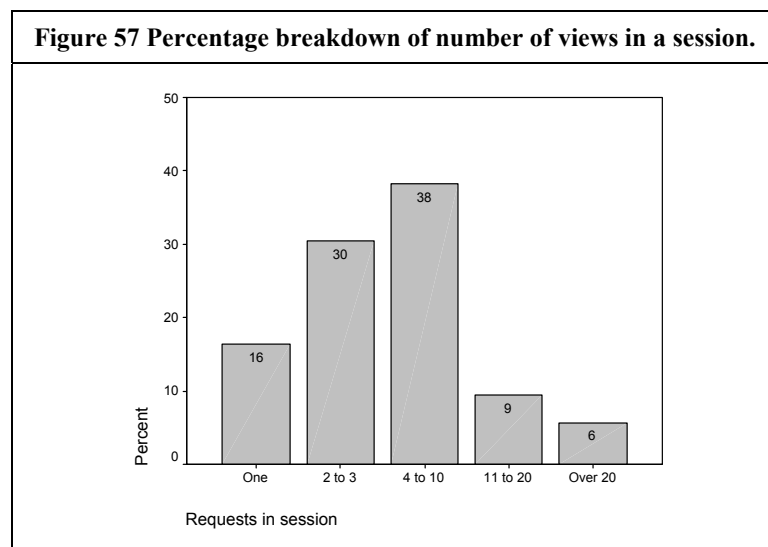
4.1.9.5 By whether respondents thought it is more important to publish in a prestigious general journal, than a MORE appropriate specialised journal

Figure 56 shows that those using Economics (3.6) and Chemistry (3.4) journals were more likely to agree with this statement, while those accessing Computer Science (2.6), Physics (2.7) and Medicine (2.8) journals disagreed with the statement.



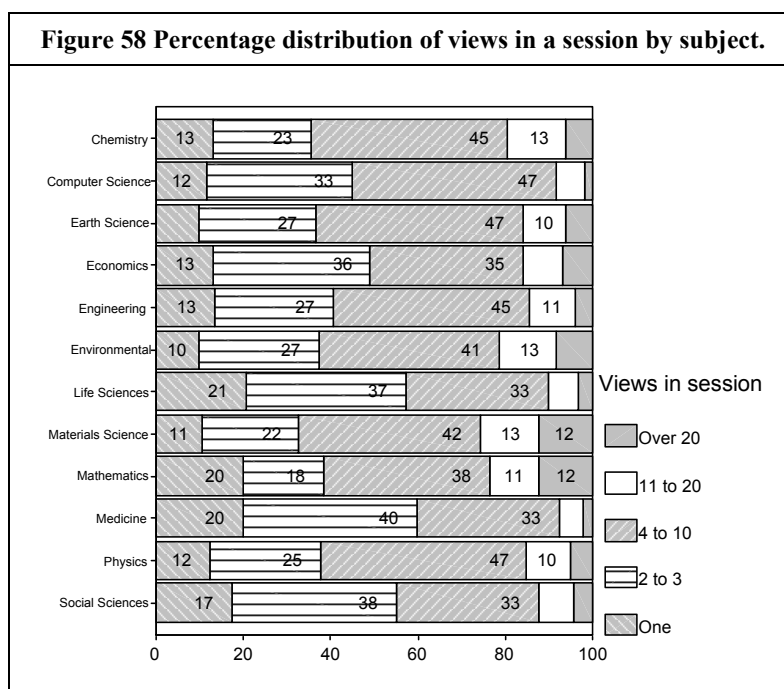
4.1.10 Items viewed in a session (site penetration)

The number of views in a session gives an idea of how deeply the user has penetrated the service, how active they were. Figure 57 gives the percentage frequency distribution of the number of items viewed in a session. Just over a third of users viewed 4 to 10 pages, 30% viewed 2 to 3, 16% just viewed one, 9% viewed 11 to 20 and 6% viewed over 20. By comparison with our studies of other digital journal libraries these are relatively high levels of penetration.



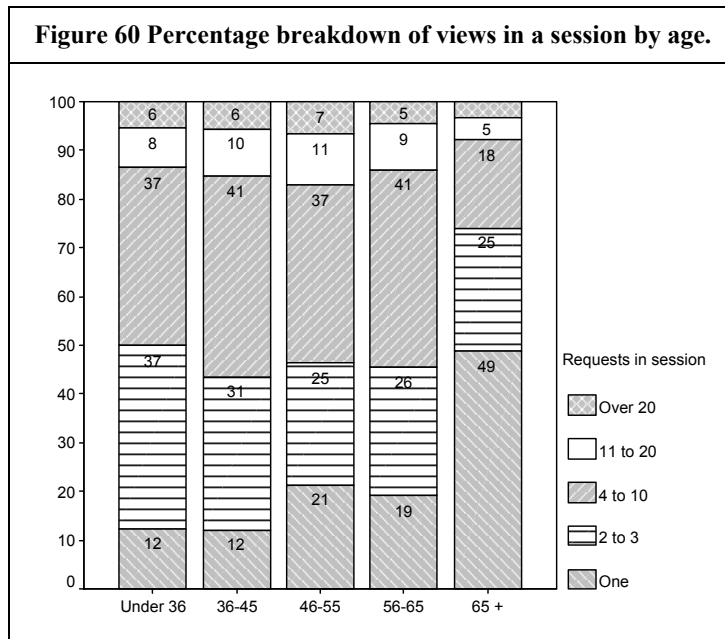
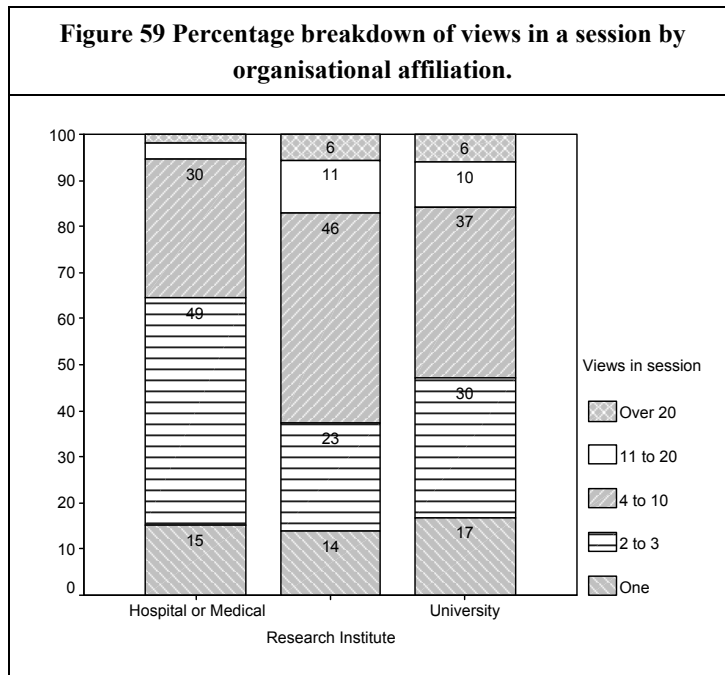
Authors as users: a deep log analysis

Life Sciences (21%), Mathematics (20%) and Medicine (20%) recorded a high proportion of sessions consisting of just one view (Figure 58). While Earth Sciences, Environmental and material sciences, mathematics and physics tended to record longer sessions (4 or more views).

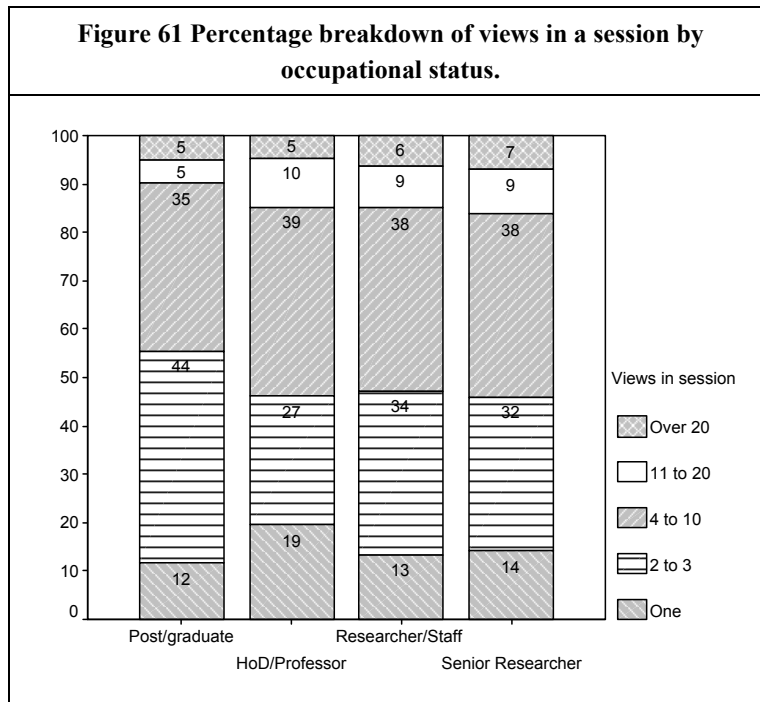


In terms of organisational affiliation (Figure 59) hospitals recorded the least number of views in a session, with 64% of sessions seeing 3 or fewer items. Research institute users conducted longer sessions and 53% of their sessions saw 4 or more items being viewed.

With regards to age (Figure 60), the number of single items viewed in a session, this marginally increases with age, from 12% for those (under 36) to about 20% for those aged between 46 and 65 rising to 49% for those aged over 65. However, users aged between 36 and 65 were likely to have long sessions and 56% conducted sessions where more than 4 items were viewed.

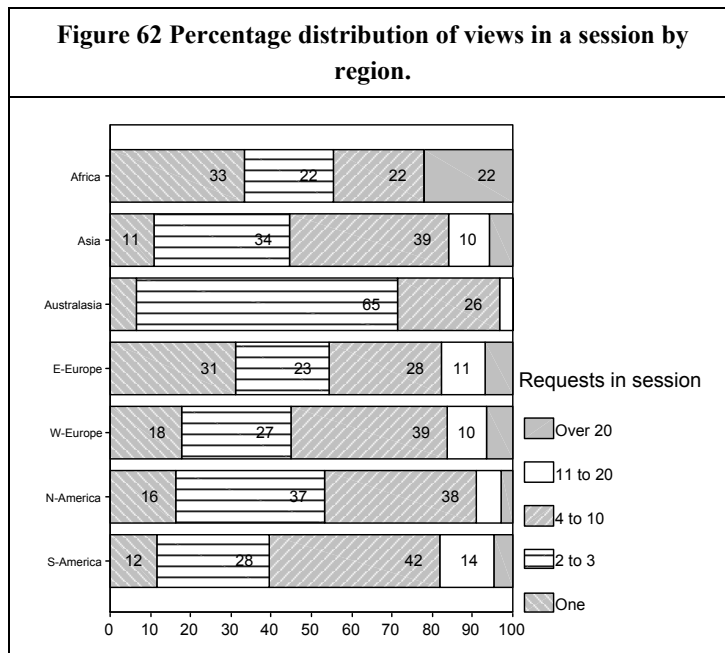


In terms of occupational status (Figure 61) there was a greater likelihood of Heads of Dept./Professors conducting sessions in which only one item was viewed, 19% of sessions were of this type compared to 13% for research staff.

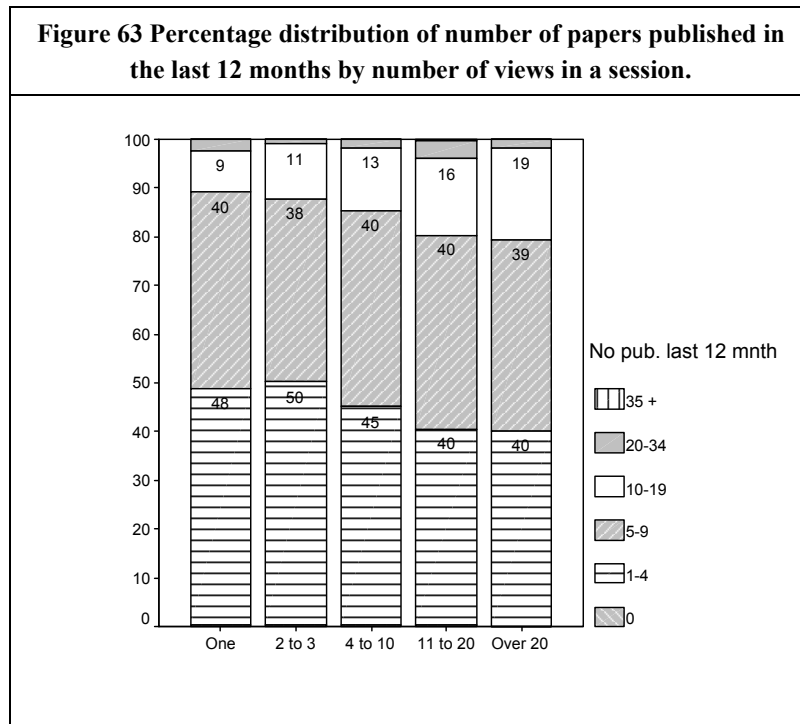


Asian and South American users recorded sessions with the greatest number of views - respectively 56% and 60% of sessions viewed 4 or more pages. African and Eastern European users were most likely to view only one item in a session.

Asian and South American users recorded sessions with the greatest number of views - respectively 56% and 60% of sessions viewed 4 or more pages (Figure 62). African and Eastern European users were most likely to view only one item in a session.



Users viewing most items in a session were more likely to have published articles in the last 12 months: 60% of those viewing 11 or more items had published 5 or more papers, compared to about 50% of those viewing 3 or less items (Figure 63).

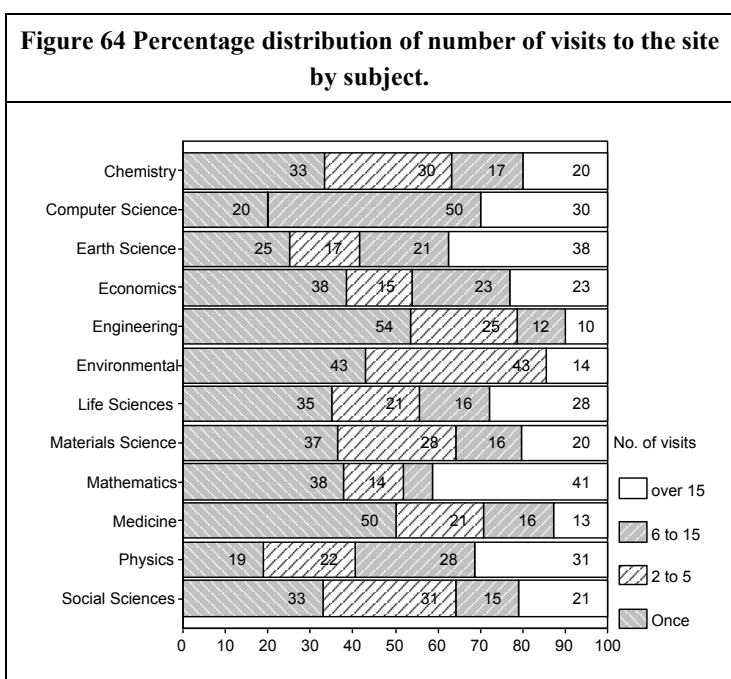


4.1.11 Return visits

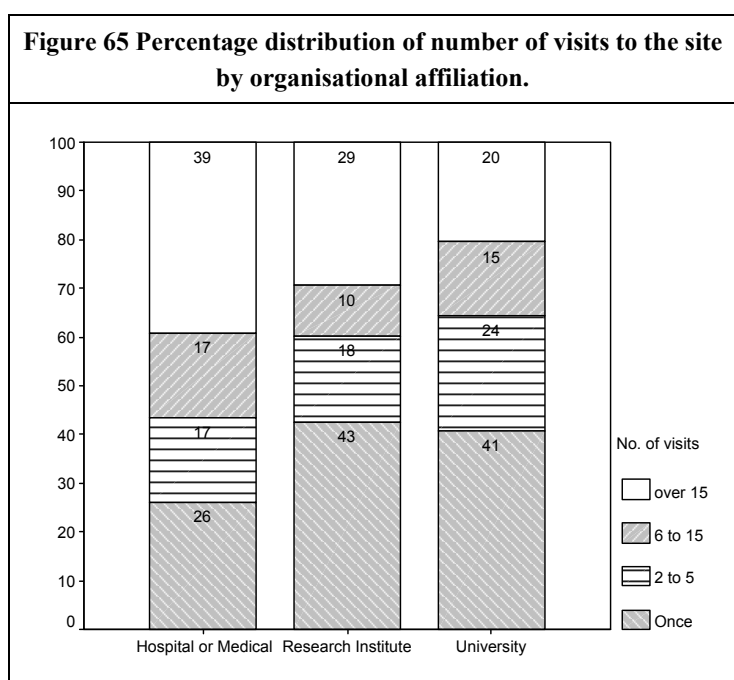
The number of times someone returns to a site to search is plainly a key metric which tells us something about site loyalty and satisfaction, arguably this is a more powerful metric than downloads. Coming back to a site appears to constitute conscious and directed use - as good an approximation of this as you are likely to get from web logs.

The metric is estimated over the five month period 1st January 2004 to 31st May 2005. Forty percent just visited once, 24% visited 2 to 5 times, 15% visited 6 to 15 and 21% of this sub sample visited over 15 times.

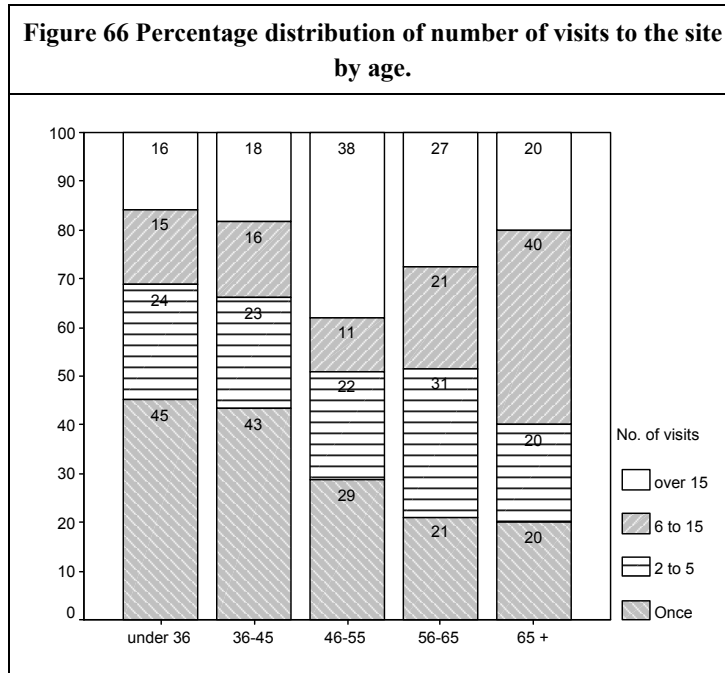
In terms of subject groupings, Business studies, Computer sciences (80%) and Physics (19%) recorded high percentages of repeat visits (these subjects also recorded high percentage of current awareness users). Medicine (50%) and Engineering (54%) recorded higher percentages of users just visiting once (Figure 64).



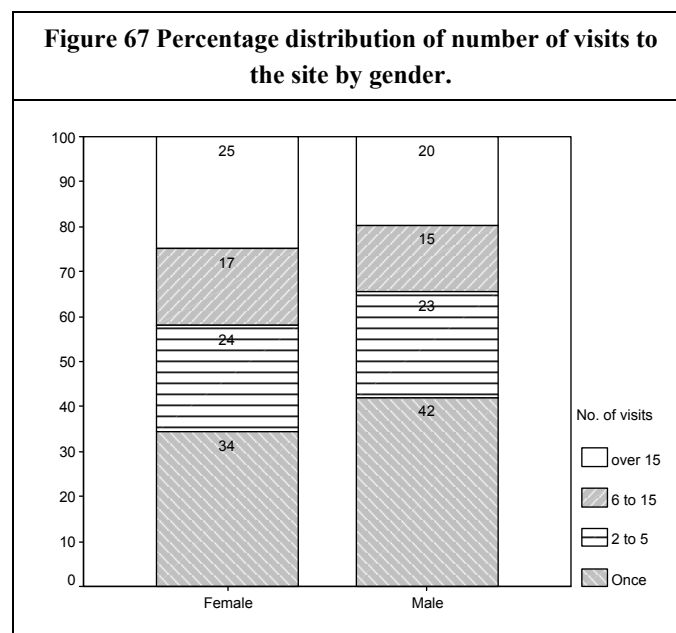
Hospitals recorded the lowest number of users visiting just once (Figure 65).



In terms of age, the likelihood that a user will repeat visit increases with age. While about 80% of those aged over 56 visited two or more times this was only true of about 55% of users aged 45 or younger (Figure 66). We have found elsewhere that the younger users were promiscuous in their information habit.

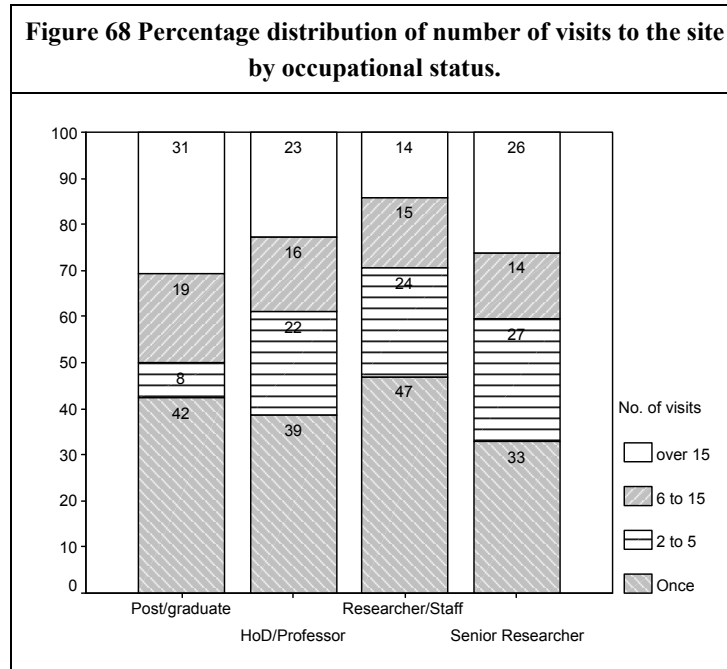


Women were more likely to return to the site compared to men; 66% visited 2 or more times compared to 58% for men (Figure 67).

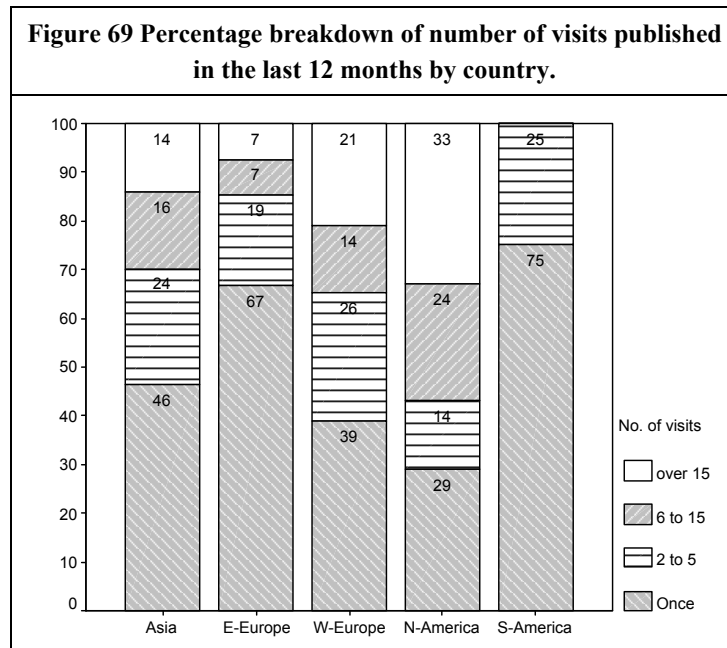


Authors as users: a deep log analysis

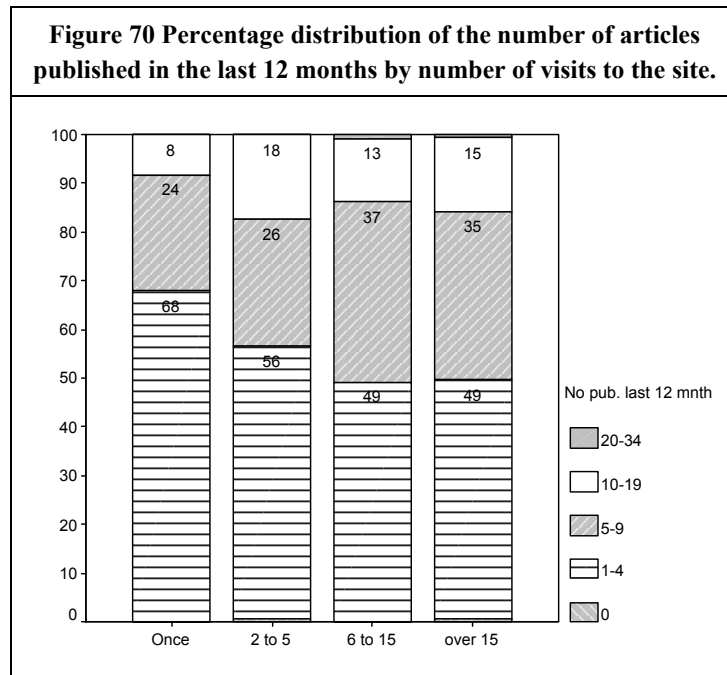
Senior researchers seemed most likely to return to the site, 66% did so as compared to 53% of research staff and 61% of heads of department/professors (Figure 68).



In terms of geographical location, users based in North America were more likely to be repeat users, 71% visited more than once while users based in South America were least likely, 75% just visited once (Figure 69).

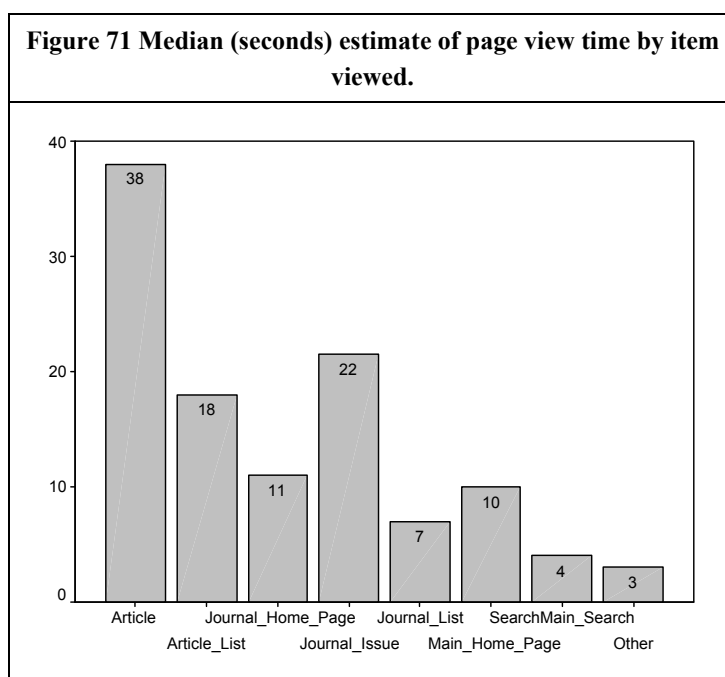


In terms of the number of publications published in the last 12 months, those people visiting more regularly tended to publish more. About half of those visiting 6 or more times had published five or more papers in the past year compared to about 32% who had done so who had just visited once (Figure 70).

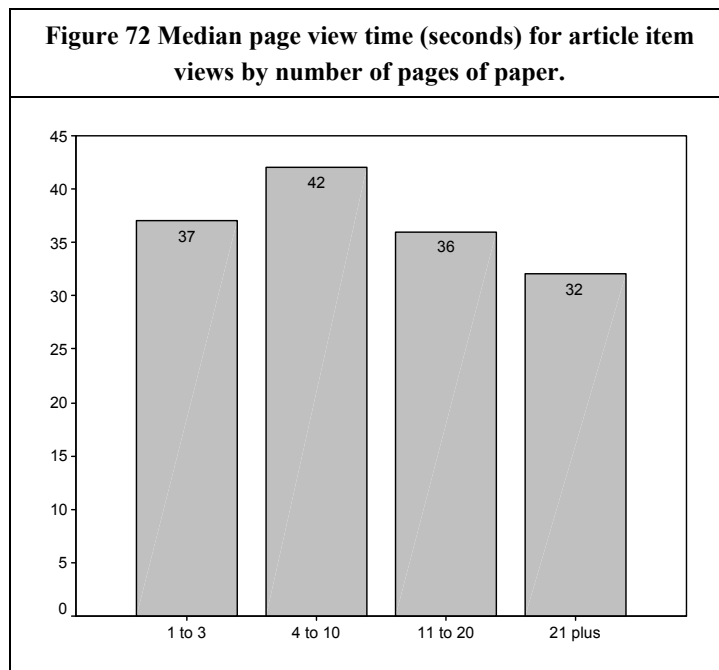


4.1.12 Time spent online

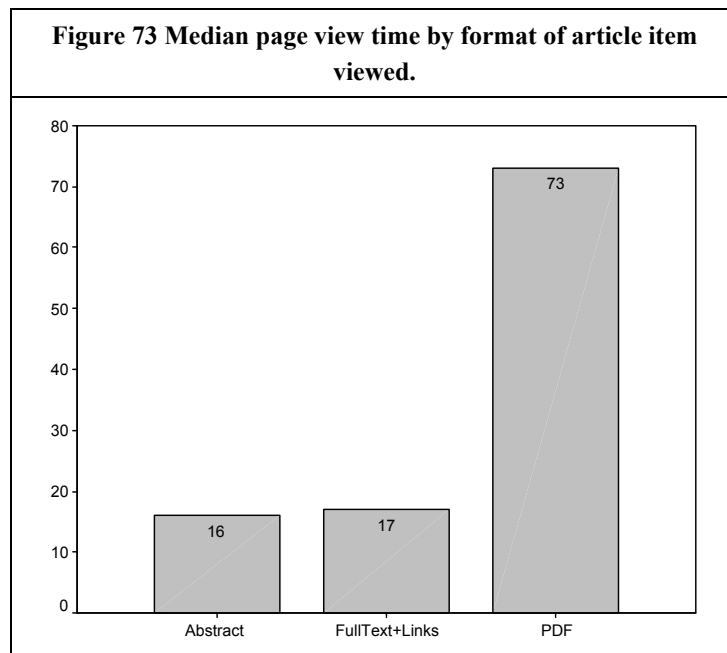
In terms of view time (Figure 71) articles took the longest to view - an average median time of about 38 seconds (which clearly suggests that people were not reading online), Journal issue pages were viewed for about 22 seconds, article lists about 18 seconds and the journal home page, Journal list and Main home page for about 10 seconds. Surprisingly the average view of the main search screen was just 4 seconds. A sign of people in a hurry, perhaps? Maybe flagging a need for ScienceDirect to capture and direct researchers to content or alternatively it's just a case of familiarity?



The second time analysis examines the influence of article length on the time taken to view an article (Figure 72). The greatest amount of average (median) online time (42 seconds) was spent on papers 4 to 10 pages long and the least amount of time (32 seconds) on those papers 21 pages or more in length. This suggests that people spend more time reading shorter articles online, again something that might have been expected. What this might mean though, as a consequence, is that shorter articles are more likely to be read than longer ones, and it follows maybe, more likely to be cited. Many of the articles downloaded to be read at another time never see the light of day again. Short articles have a higher digital visibility and we know that visibility is a major factor in consumption. However, we have to bear in mind that even 42 seconds is not enough time to read even a very short article.



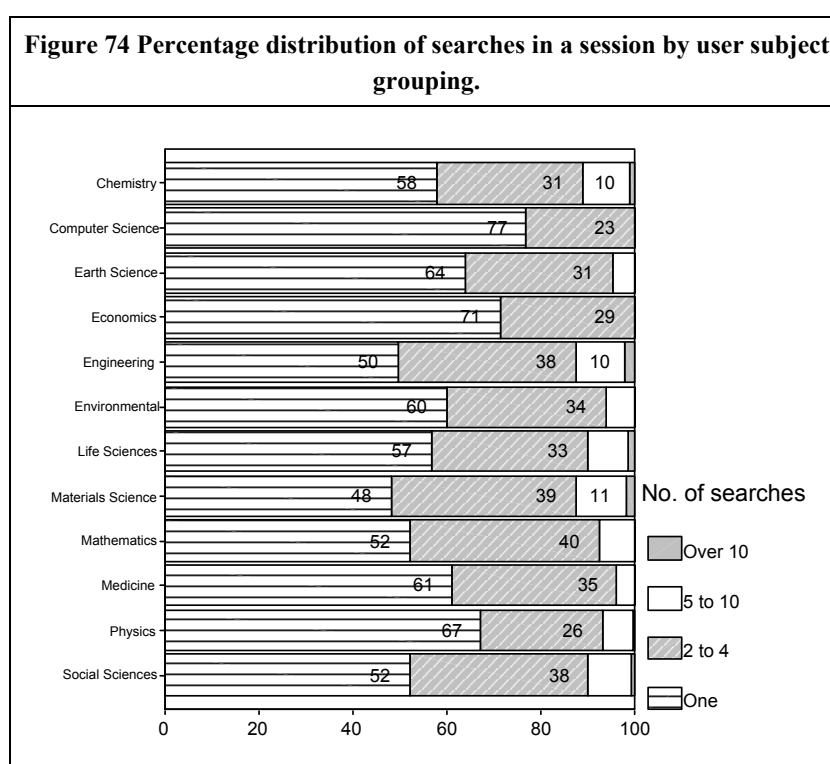
In terms of page view time (Figure 73), unsurprisingly, PDF items took longest to download and view (73 seconds), while abstract and reference screens took the shortest time to view (12 seconds). Interestingly the difference between the time taken to view an ordinary abstract and a full text HTML document is almost identical.



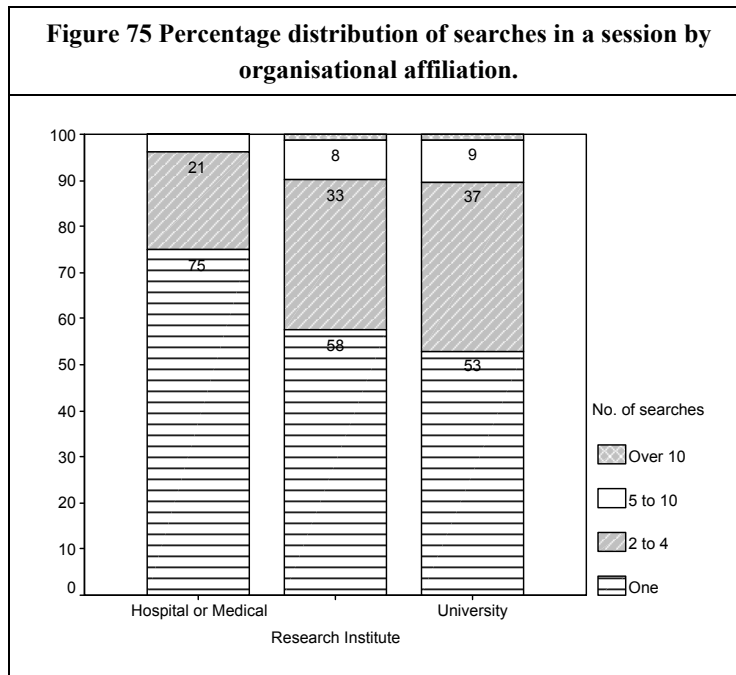
4.2. Searching and navigating

4.2.1 Number of searches in a session.

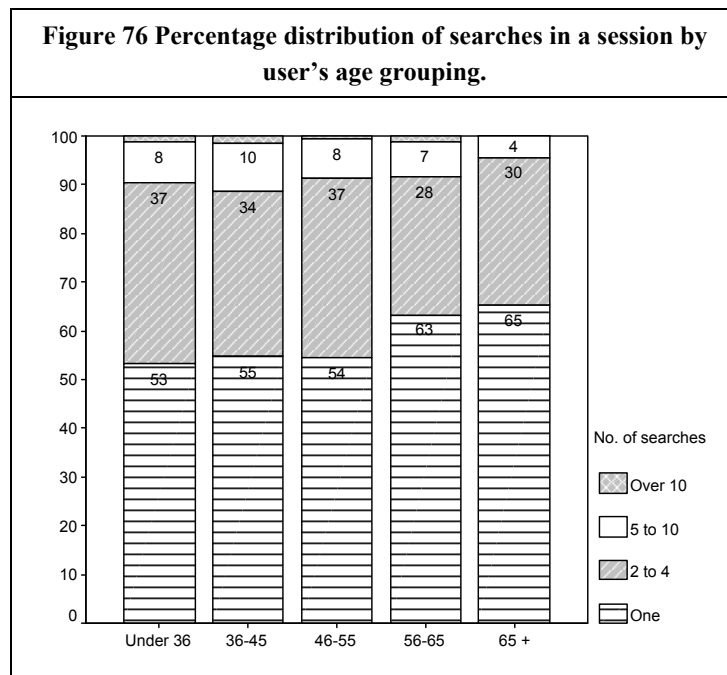
Of those sessions where a search was undertaken half saw just one search conducted, 35% saw 2 to 4 searches being conducted, 9% 5 to 10 and 1% over 10 searches. In terms of subject, users from the Computer sciences (77%), Economics (71%) and Physics (67%) were most likely just to search once in a session. Users from Material science (52%) and Engineering (50%) were more likely to make repeated searches (more than 2) in a session (Figure 74).



Academics recorded the greatest number of searches in a session (Figure 75), 47% of sessions saw two or more searches being conducted, while hospital staff recorded just 25% of such search sessions.

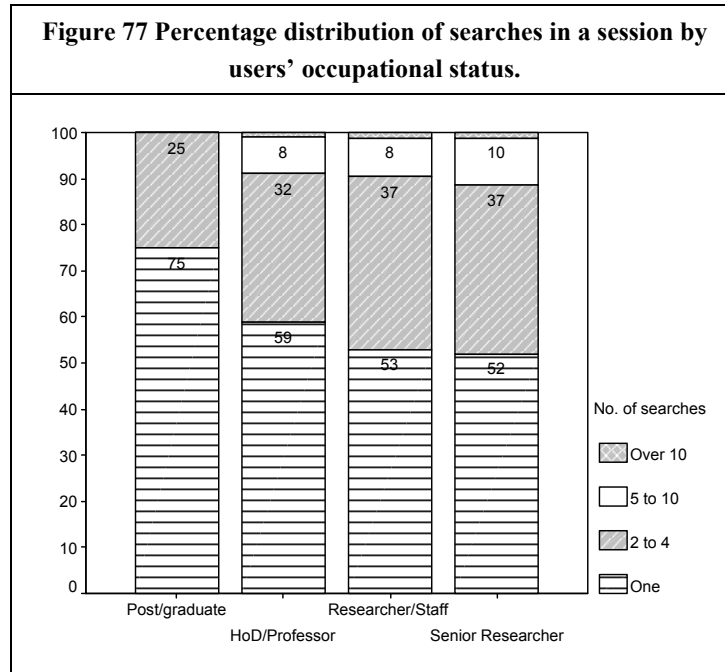


In terms age (Figure 76) the likelihood of recording sessions where only one search was executed increased with age. While 53% of those aged below 36 just made a single search, this was true of 63% of those aged over 56.

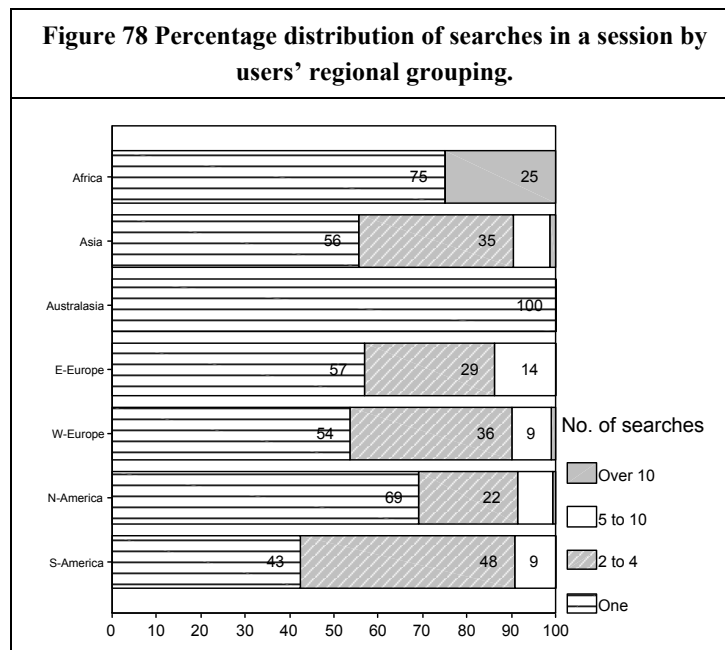


Authors as users: a deep log analysis

Students were most likely to conduct one search in a session, 75% (of those who did a search) did so, followed by Heads of Department/professors (59%). Figure 77 refers. Research staff were most likely to complete 2 or more searches in a session, about 47% did so.



In terms of region (Figure 78) users located in the Western Europe (45%) and South America (47%) were most likely to complete 2 or more searches in a session, while North American (69%) and Australasia (100) users were most likely just to search just once.

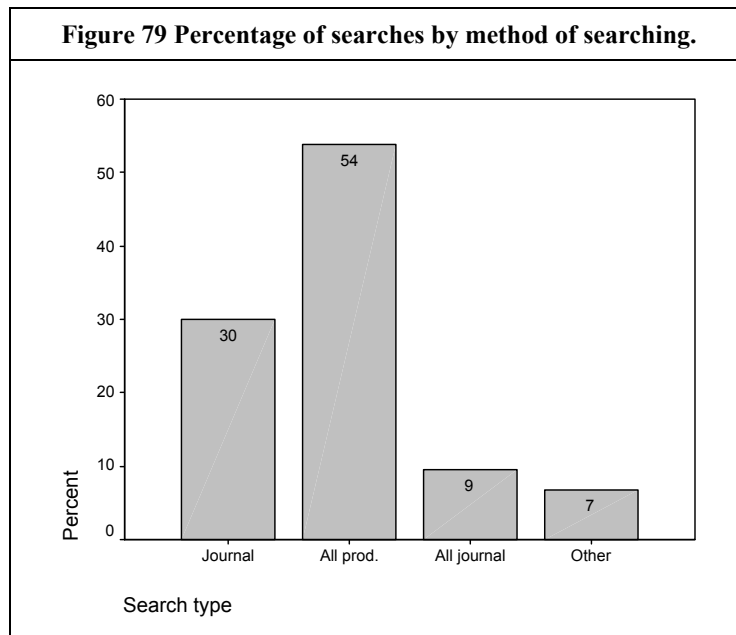


4.2.2 Search approach adopted

Users have a number of means of directly searching for content. The available search type options were:

- allbooks** - search within all book series, handbook series and reference works
- allbs** - search within all book series
- allhs** - search within all handbook series
- allinprod** - search within all full-text sources
- alljrnl** - search within all journals
- allrefw** - search within all reference works
- thisaip** - search with the article-in-press section of a journal
- thisbook** - search within a selected book (book series, handbook series or reference work)
- thisiss** - search within a selected volume/issue of a journal
- thisjrnl** - search within a selected journal
- thisrefw** - search within a selected reference work

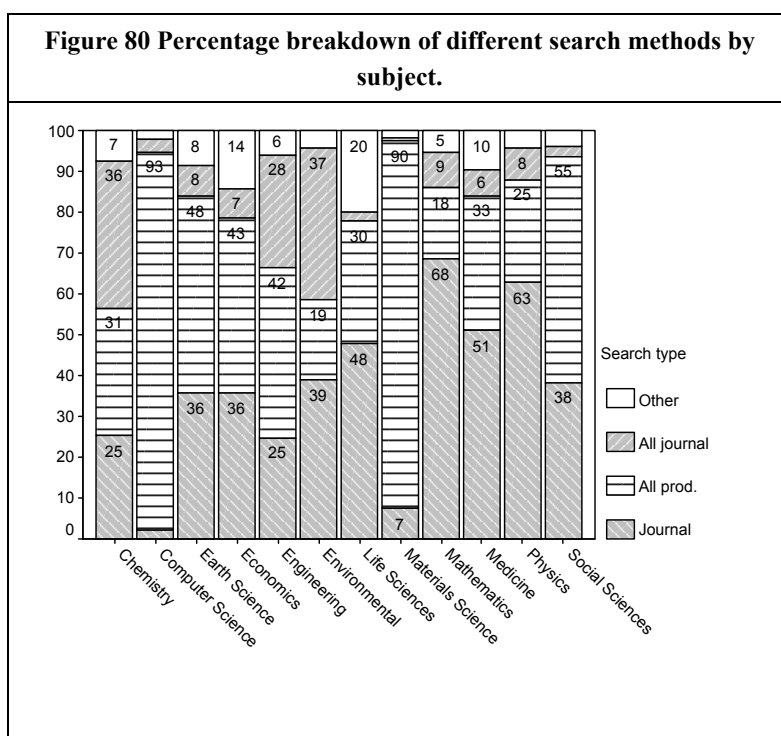
It was decided to group options into three groups: AllProd, Journal and AllJrnl. AllProd covers searches within all full-text sources, Journal covers searches within a selected journal and Alljrnl covers searches within all journals. Figure 79 gives the percentage frequency distribution of use for each. The broadest search of them all (All prod.), searching all full text sources, was the most popular accounting for 54% of searches. Just under a third (30%) were searches on selected journals (Journal), while 9% of searches used All journals.



Authors as users: a deep log analysis

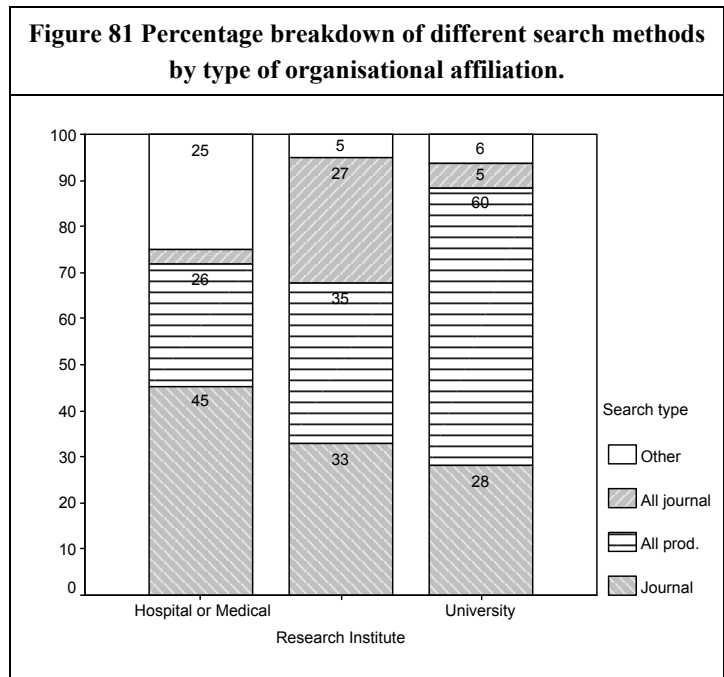
4.2.2.1 By subject background

Mathematics (68%), Physics (63%) and Medicine (51%) tended to search within a journal (Journal), a narrow and known search. It is of interest to note that both Mathematicians and physicists were also most likely just to view journals within their subject area (Figure 80). And the finding here that suggests these two subjects are well defined. Material Science (98%) and Computer Science (93%) tend to prefer searching all the full text of all sources (All prod). Chemistry (36%) and Environmental Science (37%) were proportionately more likely to use the all journal search.



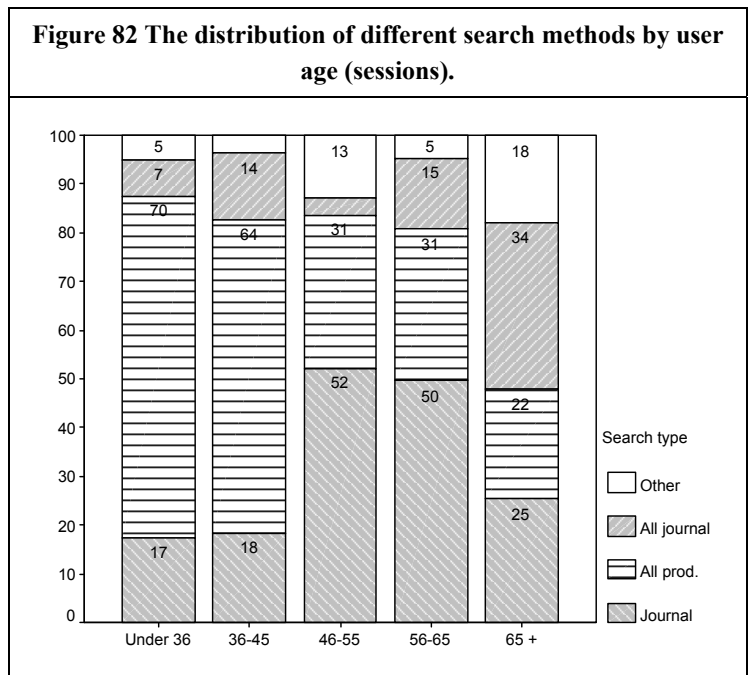
4.2.2.2 By the type of organisation

In terms of organisational affiliation, respondents working for hospitals (45%) tended to search within a journal (Figure 81). Academics (66%) made the most use of the full text documents (All prod) search. Research institutes tended to make high (27%) relative use of All journal searching.



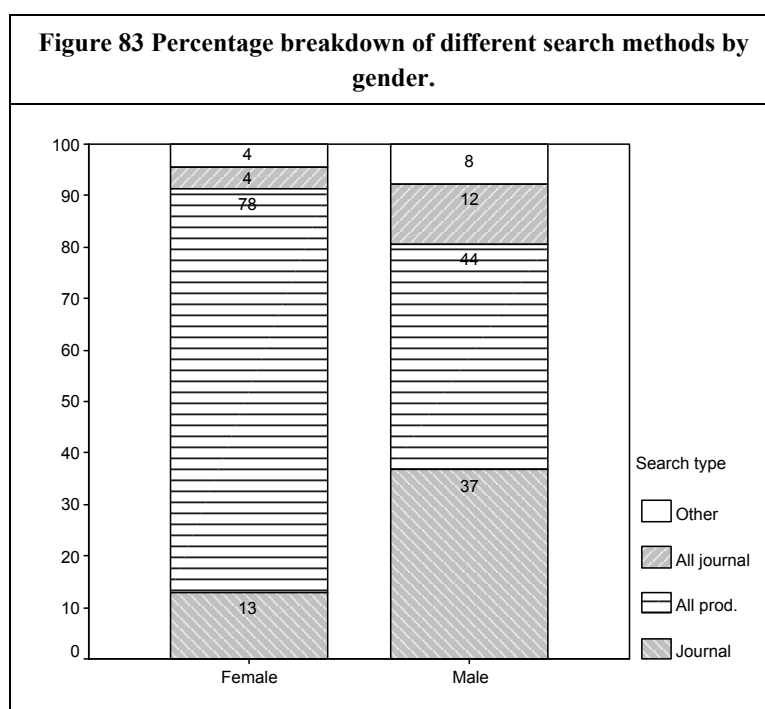
4.2.2.3 By age

The greatest use of searching within a journal was made by 46 to 65 year olds: over half (52%) of all searches were conducted that way. Figure 82 relates. Younger users tended to use search all full text (All prod) and about two thirds (70% and 64%) of searches by those aged 45 and under used this option. The use of “all journal” searches tended to increase with age, after the age of 46.



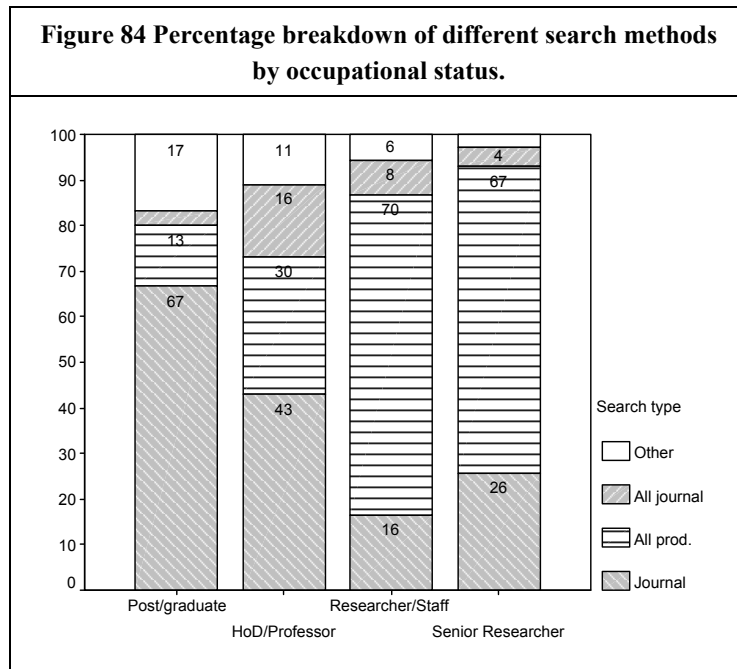
4.2.2.4 By gender

Big and surprising gender differences were discovered here. Women favoured the broad All prod search - 78% by use compared to 44% for men (Figure 83). Men favoured the all journals search - 12% of use compared to 4% for women and journal only searches -, 37% compared to 13%. Again this is another of the few analyses where gender is significant and merits follow-up investigation.



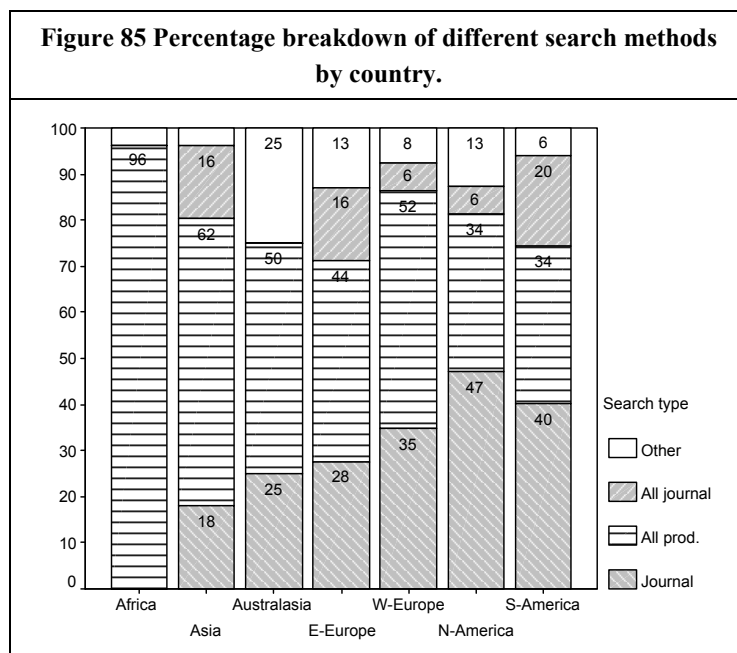
4.2.2.5 By occupational status

Again there are marked differences here, with two thirds (67%) of students employing the specific journal search (Figure 84). Researchers (70%) and Senior researchers (67%) were more likely to search using the all text sources (All prod) option. This reinforces the result found in section 4.1.8.4.



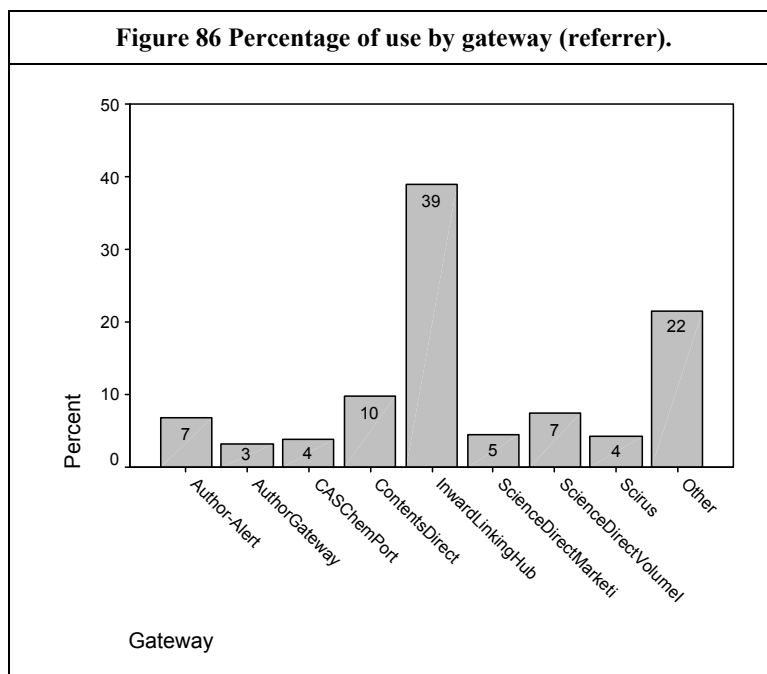
4.2.2.6 *By geographical location*

Essentially the various nations approached the system quite differently. Thus respondents from Western Europe (36%) and North America (47%) favoured the searching this journal approach (Figure 85).

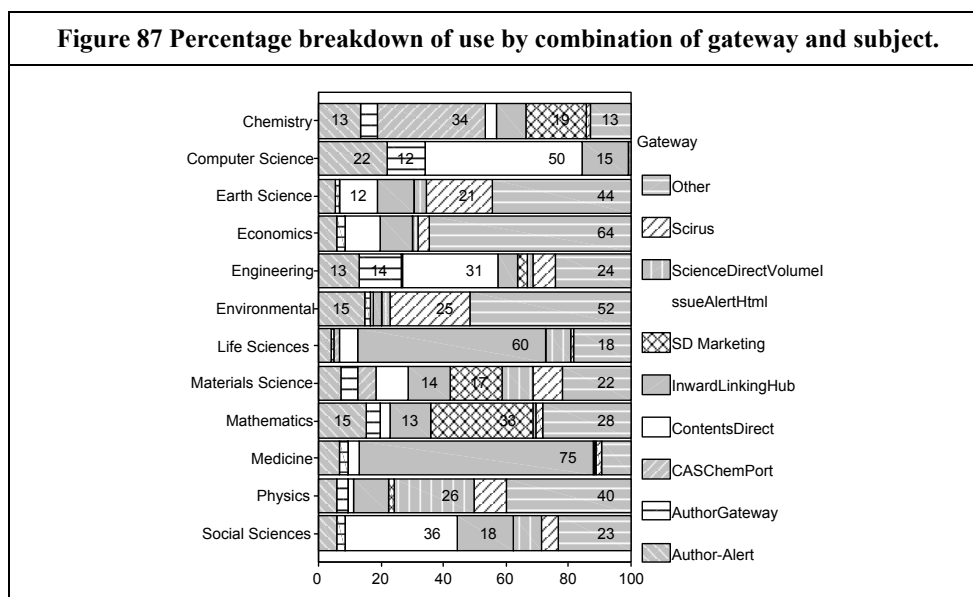


4.2.3 Gateways

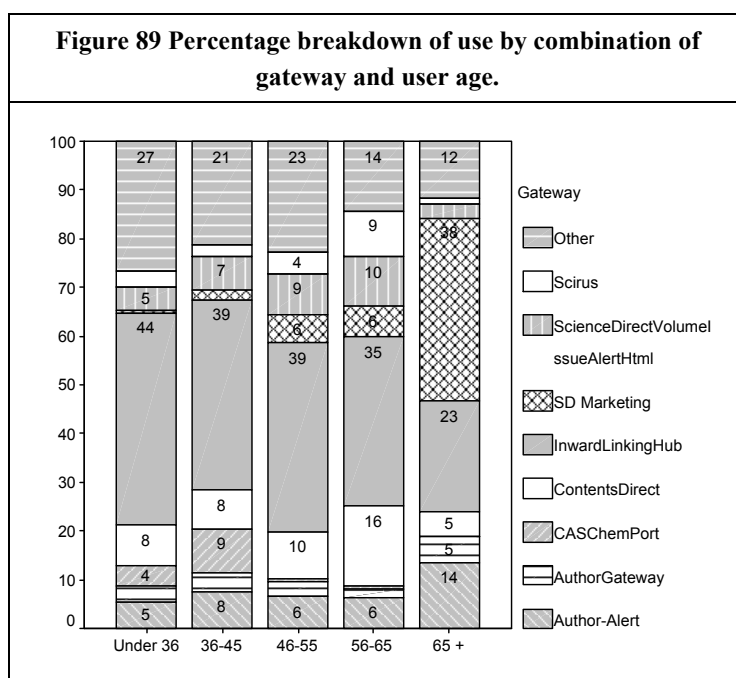
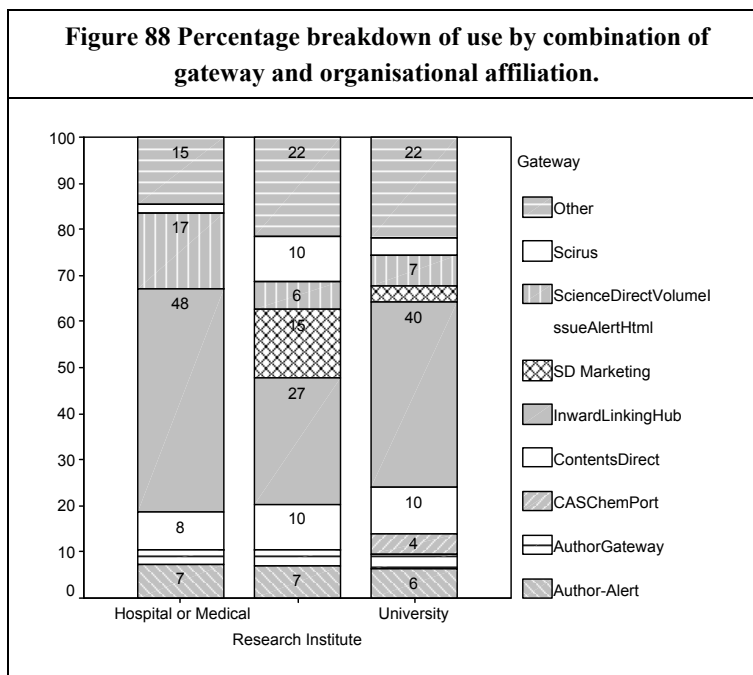
Gateways refer to external links to the site. Most Gateway users (39%) came from InwardLinkingHub (Figure 86).



InwardLinkingHub was the main Gateway for users from Medicine (75%) and Life Sciences (60%) See Figure 87. Environmental scientists made a good use of Scirus (25%), while Computer Science (50%) mainly used Contents Direct. In terms of ScienceDirect alert links this featured strongly in Physics (26%) while ScienceDirect marketing gateway links were popular in Mathematics (33%), Chemistry (19%), and Material science (17%).

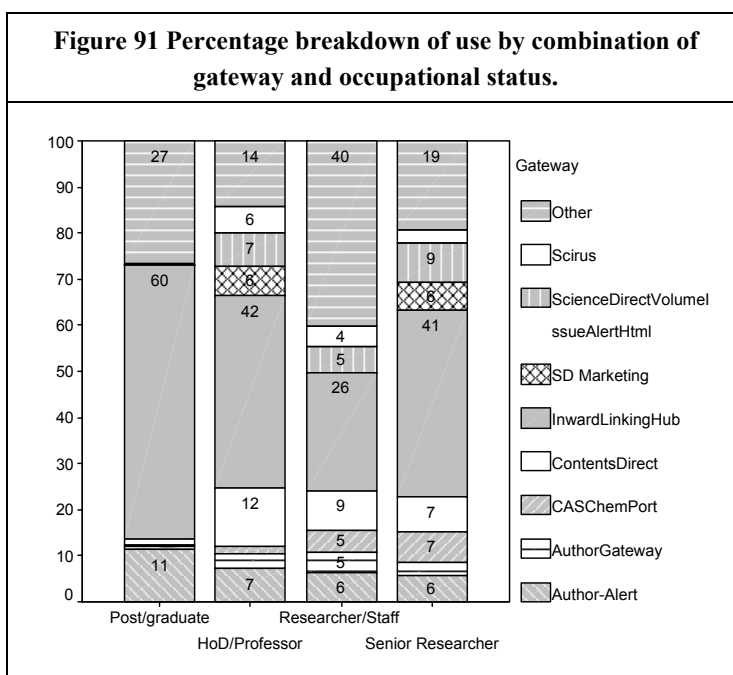
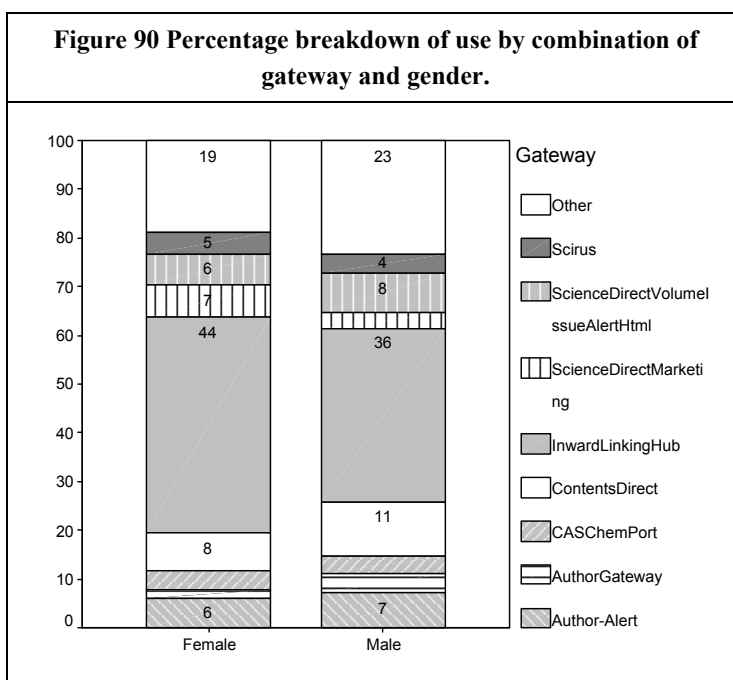


Respondents based in hospitals were more likely to come from InwardLinkHub (48%), although a relatively high percentage (17%) came in via Science Direct Alerts (Figure 88). Fifteen percent of Gateway users in industry came in via Science Direct Marketing. Gateway links via Science Direct Marketing or alerts appear more important as the respondent's age increased; this was also true of use via Scirus (Figure 89).

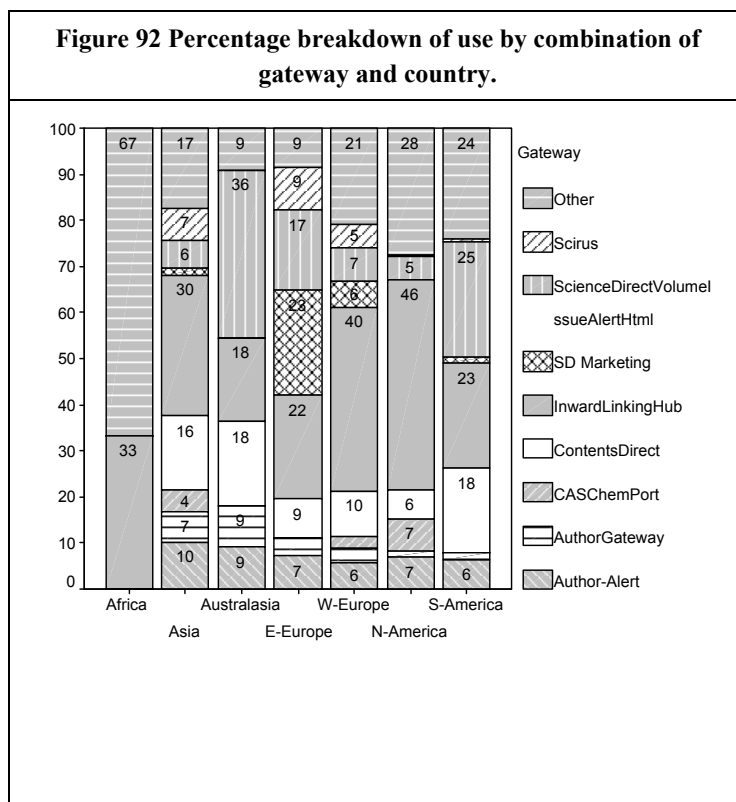


Authors as users: a deep log analysis

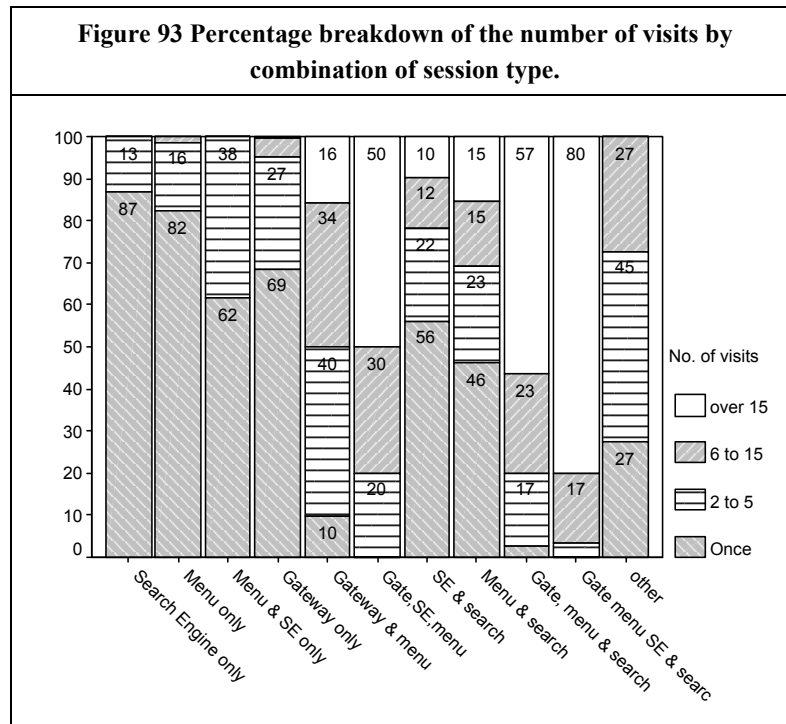
Women were significantly more likely to access the service through InwardLinkingHub; 44% did so as compared to 36% for men (Figure 90). Graduate gateway respondents mainly entered via InwardLinkingHub (60% of them did so). Figure 91 relates. Professors and senior researchers tended to use Science direct marketing or alerts, about 13% did so. Researchers/staff were least likely to use the gateway Inwardlinking hub (26%), 40% of this group used some other gateway link.



North American and Western Europe respondents favoured the Inwardlinking hub (46% and 40% respectively) and respondents in Eastern Europe (23%) used ScienceDirect Marketing links. Australasian respondents were relatively high users of Science Direct alerts (36%). See Figure 92.



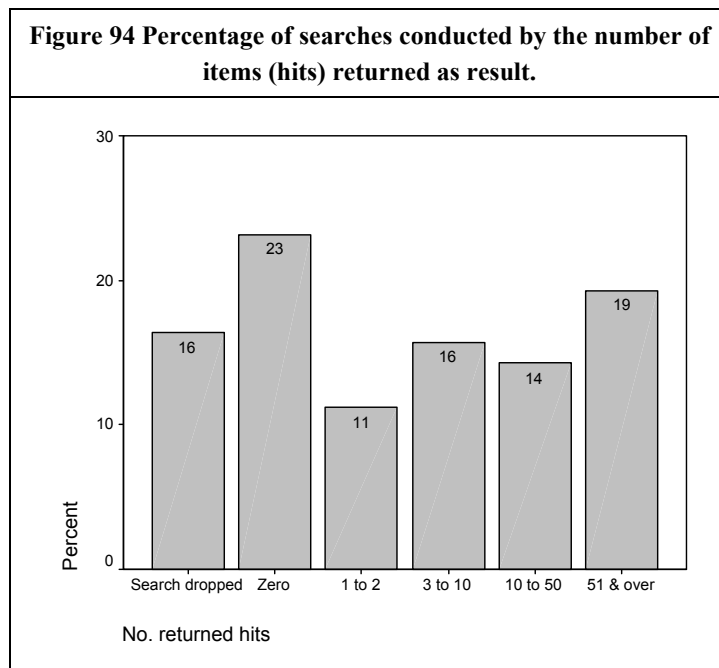
Sessions were classified into whether the user had started the session via a search engine, via a gateway, had used a menu (excluding homepage) or if the user had executed a successful search. Users, of course, might use a variety of these options for different sessions. Figure 93 looks at the variety of session types by the distribution of the number of visits for each. Those entering just via a search engine were most likely just to visit once, 87% did so. Those users just using menus or only entering the site via a Gateway again were likely just to visit once - respectively 82% and 69% did so. Those users undertaking sessions combining gateway services, menus and the search facility were the most likely to return to the site, more so than the Science Direct only user (menu & search).



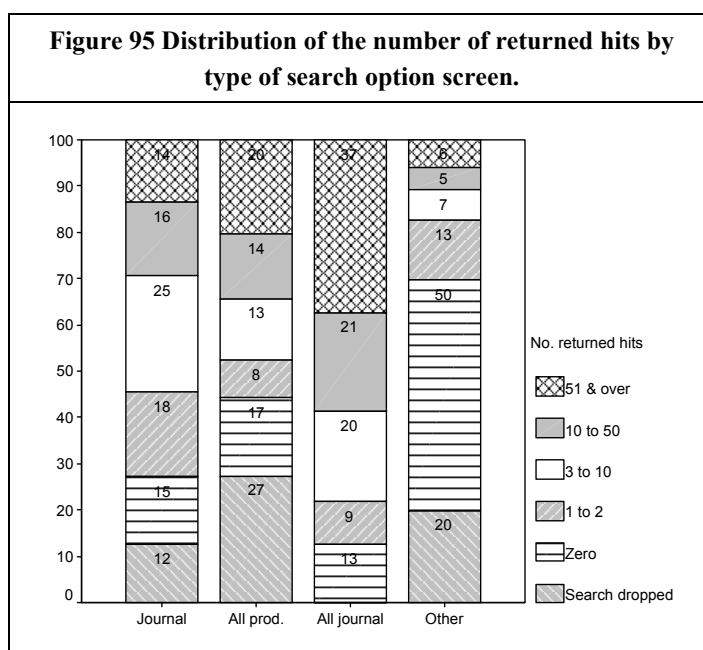
4.2.4 Number of returned hits (grouped)

The number of returned hits refers to the number of matches returned by the internal search facility in response to a query. This covers all options where the internal search facility was used. It tells us something about the effectiveness, productivity or specificity of the search, or, perhaps of the comprehensiveness of the database being searched.

About 16% of the sample abandoned their search after viewing the search screen and a further quarter (23%) recorded zero hits in response to a query (Figure 94). This represents what appears to be a very high failure or bounce rate and as a matter of priority should merit further research. Of the remaining searches, 11% obtained 1 to 2 hits, 16% obtained 3 to 10 returned hits, 14% obtained 10 to 50 and 19% obtained 51 or more. This result further supports our contention that users are not really penetrating deeply into the database.

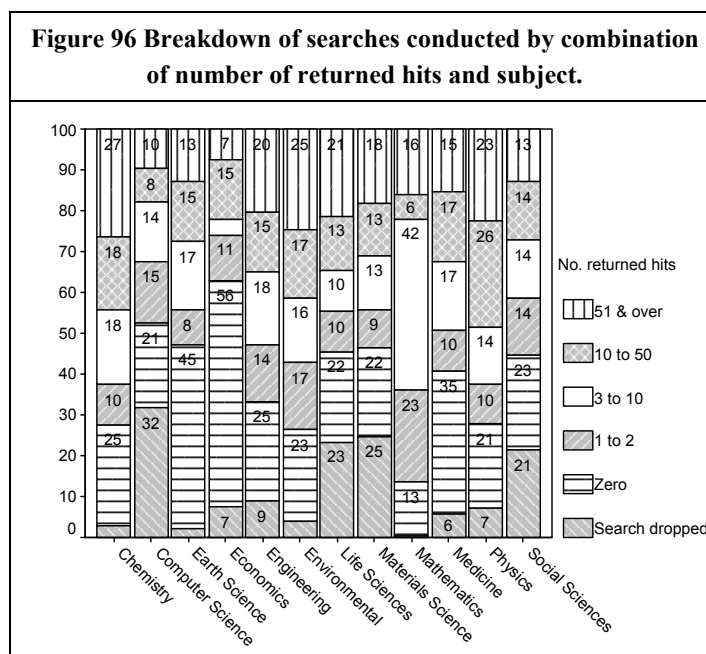


However, the number of returned hits varied dramatically depending of the type of search approach adopted (Figure 95). Searching all journals appeared to be the most successful – taking success to be the number of returned hits (but of course this might not necessarily be the case). About 60% of searches using the All journal search option were returned with 10 or more hits; this was only true 34% using the All prod option and 30% using the just this journal search. Users using the all prod option (searching all text) were less likely to complete a search – 27% of these searches were dropped, and were more likely to be returned no hits – 17% of these searches scored zero hits. However, one can see inexperienced users will be drawn to this search option.



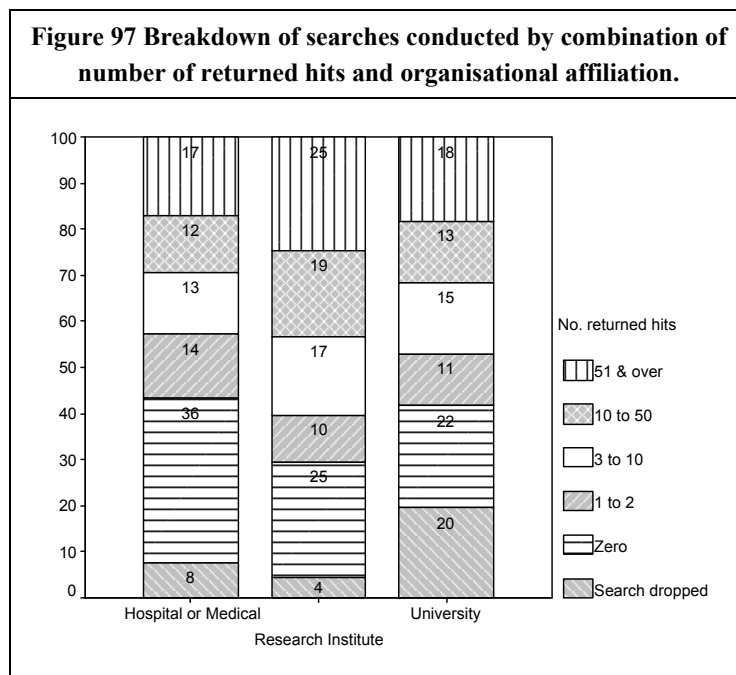
4.2.4.1 By subject background

In terms of subject (Figure 96), Computer Science (32%) and Material Science (25%) recorded high drop out rates. Those in Earth Sciences (46%) and Economics (56%) recorded high percentages of zero returns. In terms of returned hits Engineering (67%) Environmental Science (74%) and Physics (74%) appeared to have performed well, with 25% of searches by Environmental Science generating more than 50 hits.



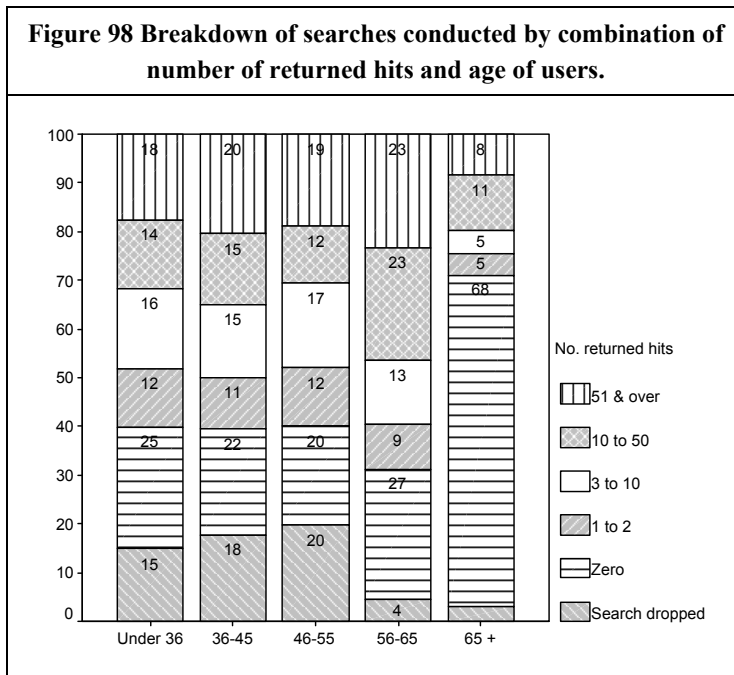
4.2.4.2 *By type of organisation*

About a fifth (20%) of respondents in universities dropped out of the search screen and a further fifth (22%) and over a third (36%) of those searching from a hospital were returned zero hits (Figure 97). Those respondents searching from research institutes were most likely to be returned hits; about 81% were successful in being returned one or more hits, which might be an indicator of their greater literature knowledge and search proficiency.



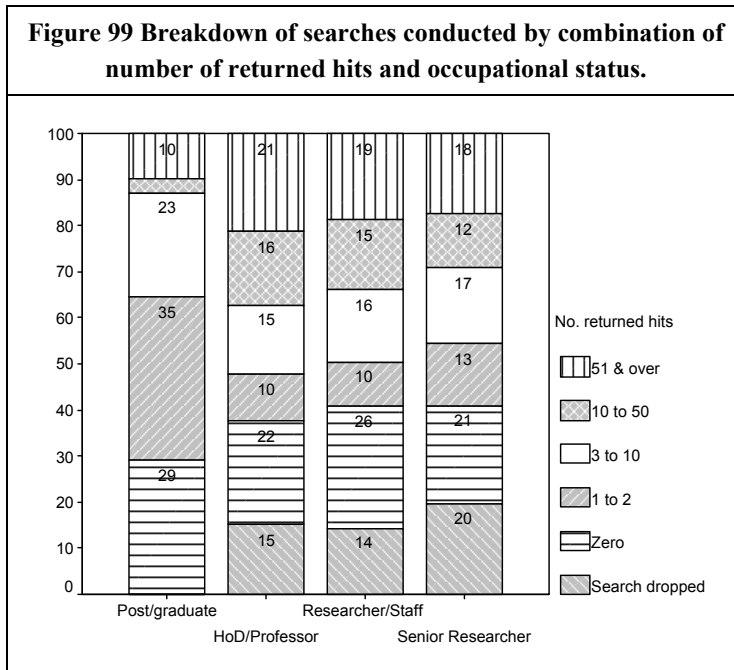
4.2.4.3 *By age*

Elderly users appeared to perform most ‘badly’. Over two-thirds (68%) of searches conducted by respondents over 65 obtained zero returns in response to a search (Figure 98). This has to be regarded as being astonishing. However, those just a little younger (those aged 56 to 65) appeared most successful in obtaining hits - about 79% had 1 or more hits in response to a search.



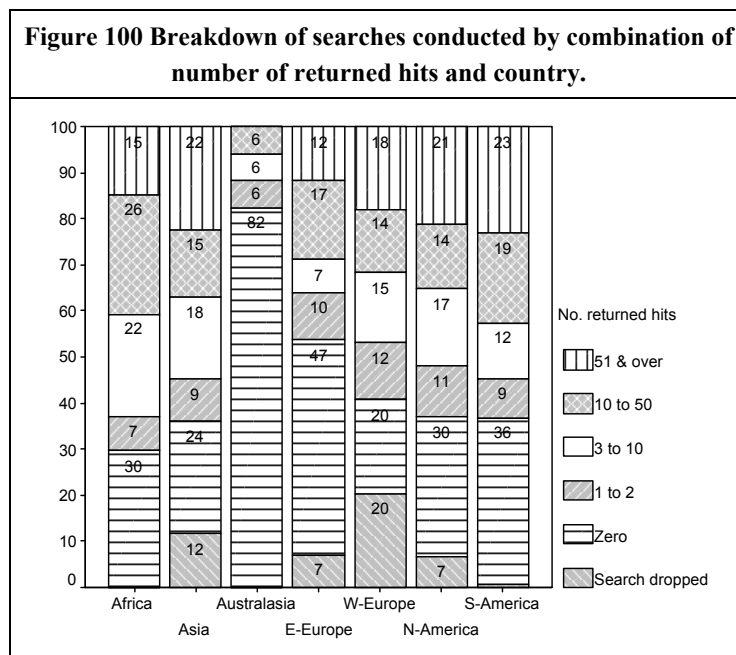
4.2.4.4 *By occupational status*

Senior researches were most likely (20%) to abandon their search (they might of course have chosen to conduct their search in another mode/screen (Figure 99). However, overall, students seemed to encounter the greatest difficulty, with about two thirds of them either dropping their search or recording zero hits. Heads of Department, marginally, appeared most able to secure returned hits, 63% of searches achieved one or more hits.



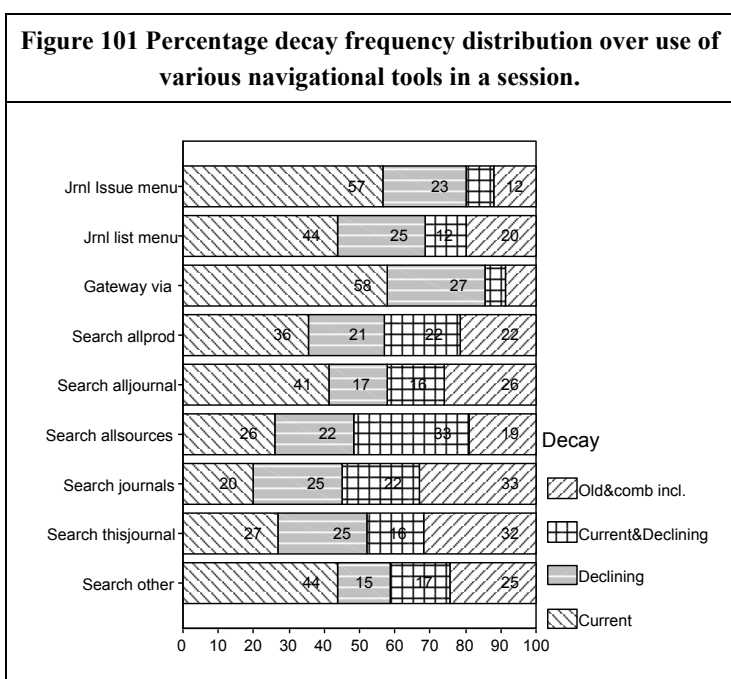
4.2.4.5 *By geographical location*

In terms of searches abandoned respondents from Western Europe (20%) rated most highly (Figure 100). Respondents based in Eastern Europe (47%) and Australasia (100%) recorded a relatively high percentage of searches resulting in zero returns. North Americans appeared the most competent or successful searchers and about 63% of their searches resulted in one or more matches.



4.2.4.6 *By age of article item viewed*

Figure 101 gives the percentage frequency of the age of material consulted by the various ways of searching the database. The use of navigational tools was not exclusive and users may have used a combination. However, there is strong evidence to suggest that those using a search option rather than a menu (browsing) option were less likely to view current material. About 57% of those using a journal issue menu will look at current material; however, this is true of 26% executing an all sources search and only 20% of those executing the search journals option.



4.3. Attitudinal data

The aim of this section is to show what could be learnt when navigational and viewing behaviour metrics, as identified in the ScienceDirect logs, were related to demographic and attitudinal data from the questionnaire survey of authors' scholarly behaviour and attitudes. In particular we wished to see what such an analysis would tell us about: a) the core functions of scholarly publishing - certification, dissemination, registration and archiving, identified by Mabe and Amin (2002); b) models of scholarly information seeking behaviour. We were also interested to see whether questionnaire responses agreed with what users actually did (as recorded in the logs).

4.3.1 Core functions

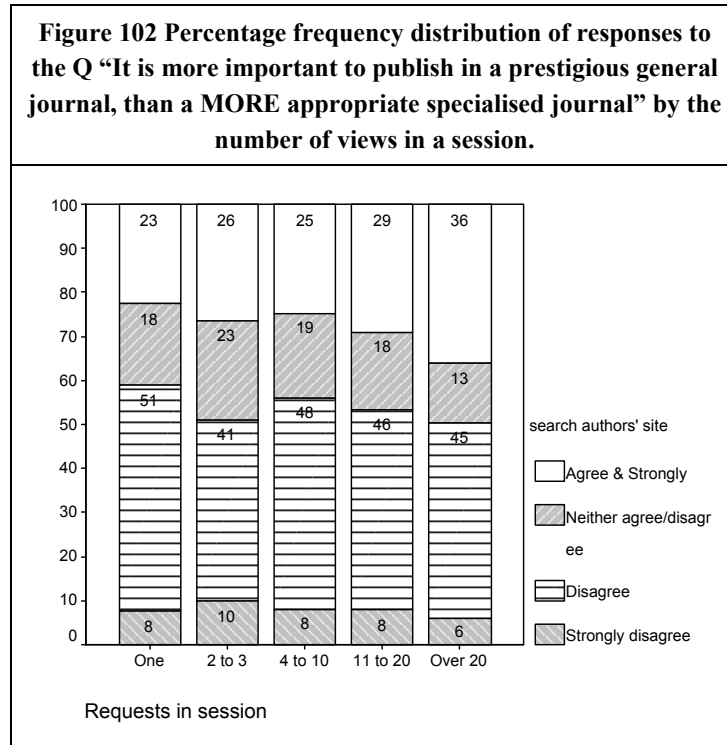
One of the purposes of the questionnaire was to provide knowledge of these functions and various questions sought to elicit information about them. Log analysis sought to relate questionnaire responses to information seeking to enrich our understanding of core functions.

4.3.1.1 CERTIFICATION

4.3.1.1.1 Where to publish?

Figure 102 gives the percentage frequency distribution of responses to the question "It is more important to publish in a prestigious general journal, than a MORE appropriate

specialised journal” by the number of views made in a session. Those viewing more items (i.e. the more active users) in a session were more likely to agree with this statement (i.e. were more concerned about status). Thirty-six percent of those viewing over 20 items in a session agreed compared to 23% who agree who just viewed one.



4.3.1.1.2 Peer review/article status

In this connection we sought to determine whether respondents who agreed with the two statements 'Readers do NOT really need refereed journals' and "The quality of an article is determined by the journal within it is published" searched differently from those that did not, with the hypothesis being that those people who disagreed with the first statement and agreed with the second statement would limit their searching to a selected group of titles. Figure 103 looks at the first statement against the number of different journals searched. It is very clear that those people that viewed a lot of journals in a session agreed that readers did not really need refereed journals. In the case of the second question (Figure 104) there was only a little evidence to support the thesis, with 6% of those viewing one journal strongly agreeing with the statement, and just 2% of those viewing five or more journals saying this.

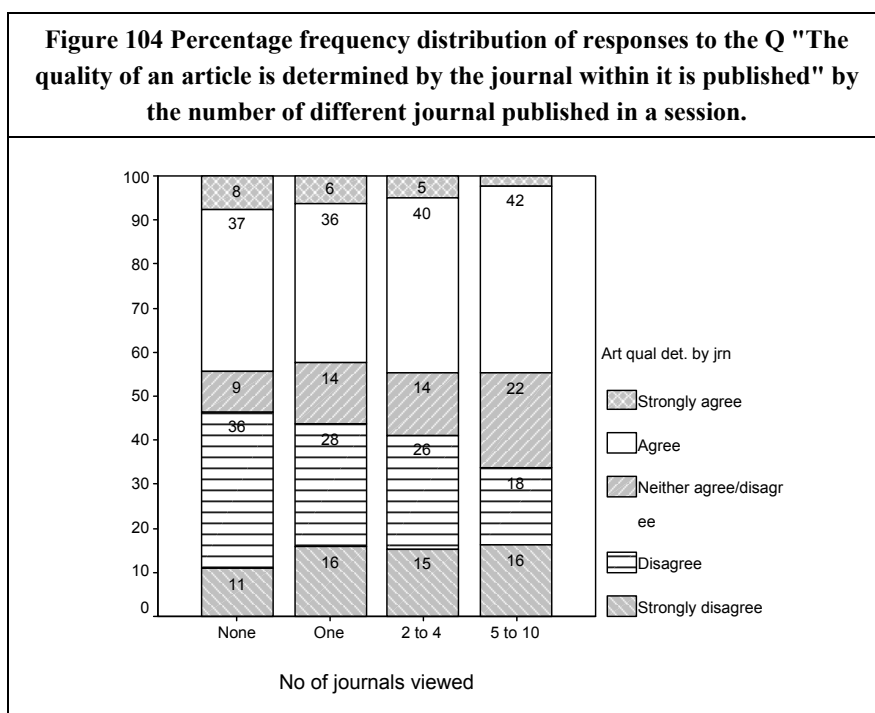
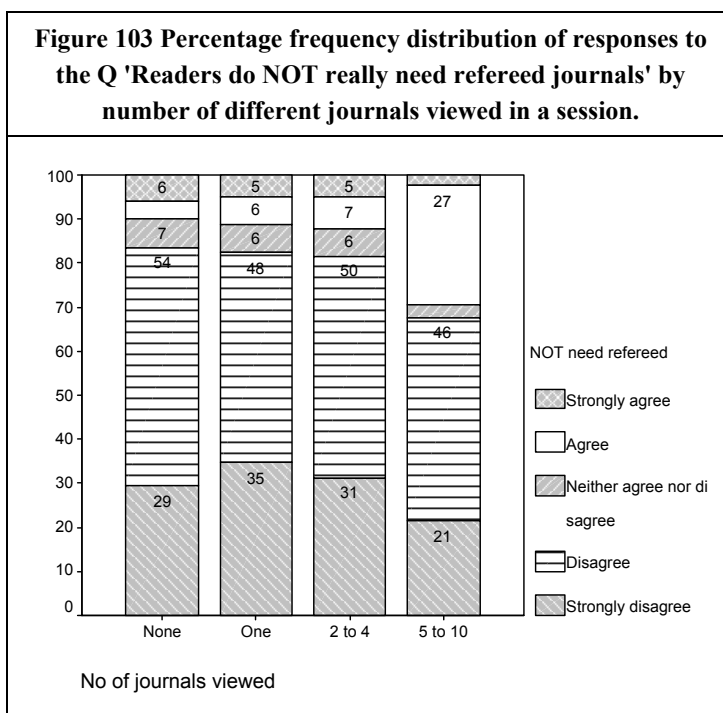
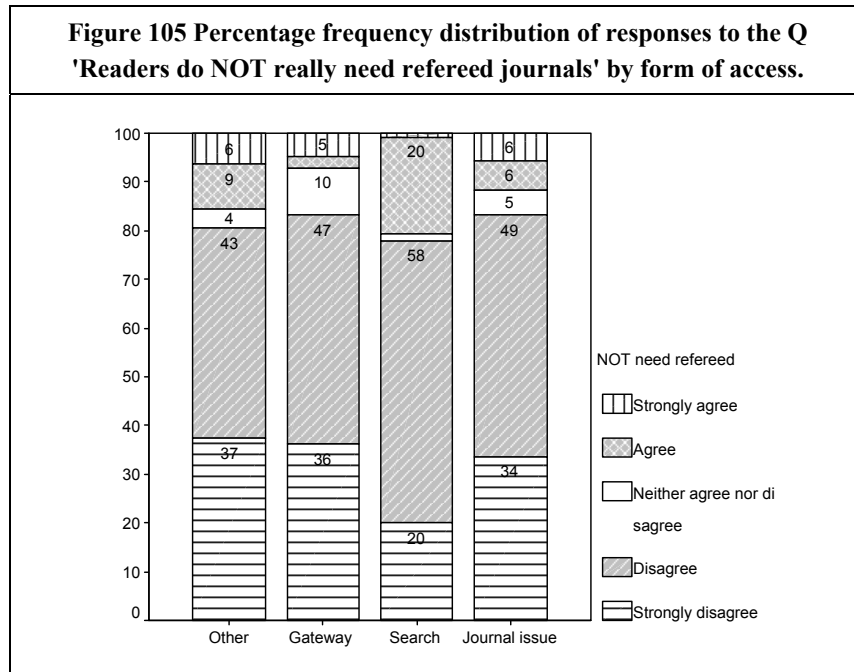
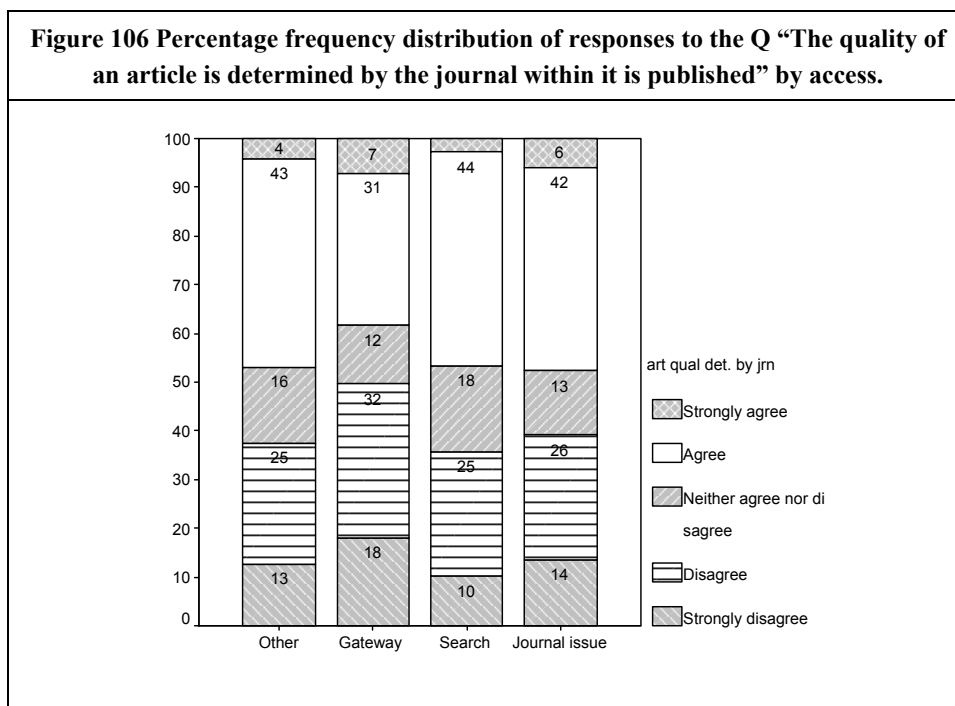


Figure 105 takes a slightly different tack by relating responses to the question 'Readers do NOT really need refereed journals' by the form of accessing ScienceDirect, with the hypothesis that those users browsing journal issues were more journal aware and thus more likely to

disagree with the question and search users less likely to. And so it proved with those 20% of those respondents using the internal search facility strongly agreeing compared to a figure of 34% for journal issue users.



A similar analysis was conducted for the question "The quality of an article is determined by the journal within it is published". Those respondents using accessing the site via a gateway were more likely to disagree and less likely to agree with the statement (Figure 106).



4.3.1.2 DISSEMINATION

4.3.1.2.1 Browsing journals from home

With the greater opportunities to search journals from the comfort of the home we wanted to discover whether people preferring to search from home were different in their viewing behaviour in any way from those preferring to search from work. Two aspects of information seeking were considered: 1) the number of views in a session; 2) the age of the material viewed. Figure 107 gives the percentage frequency distribution of responses to the question “I prefer to do my e-journal browsing at home rather than at work” by the number of views in a session. Those viewing more pages in a session were more likely to disagree with the statement. Seventy-one percent of respondents viewing over 20 views disagreed or disagreed strongly, compared to 53% of those just viewing 2 to 3 views. It appears that those people who are characterised by having heavier or busier sessions tended not to do their searching from home.

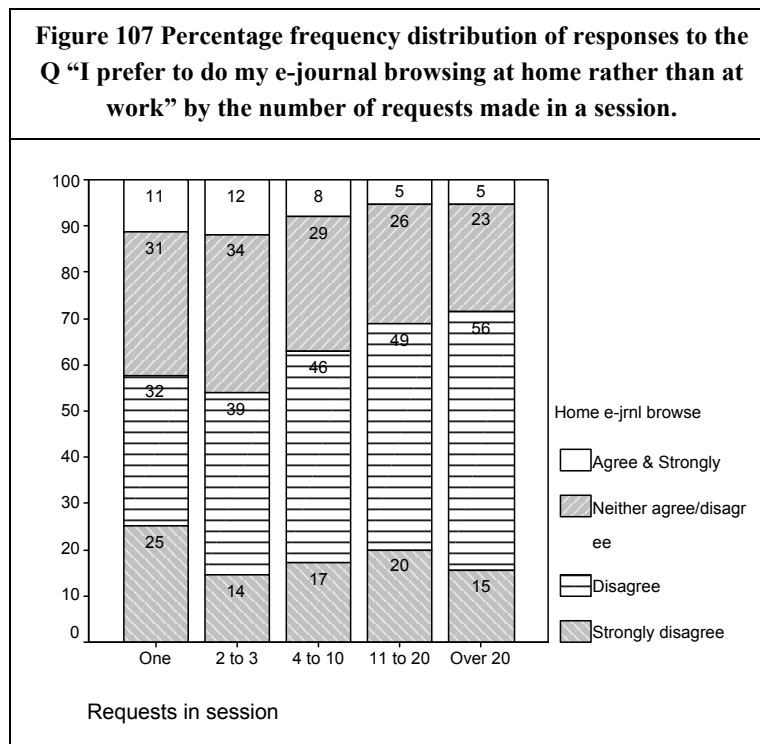
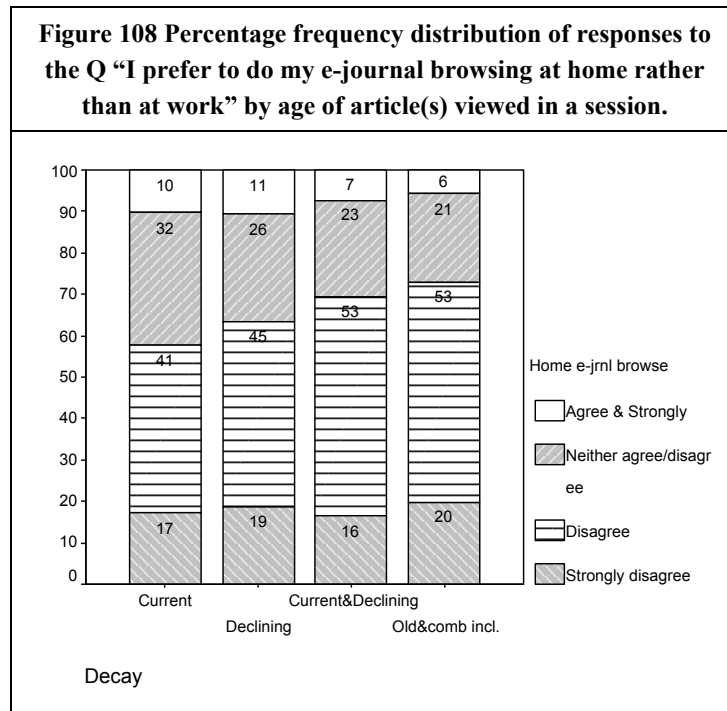


Figure 108 shows the percentage frequency distribution of responses to the same statement by the age of article(s) viewed in a session. Those viewing older material were more likely to disagree with this statement. About three quarters of those viewing items that included a view to old material disagreed compared to 58% of those viewing just current material who

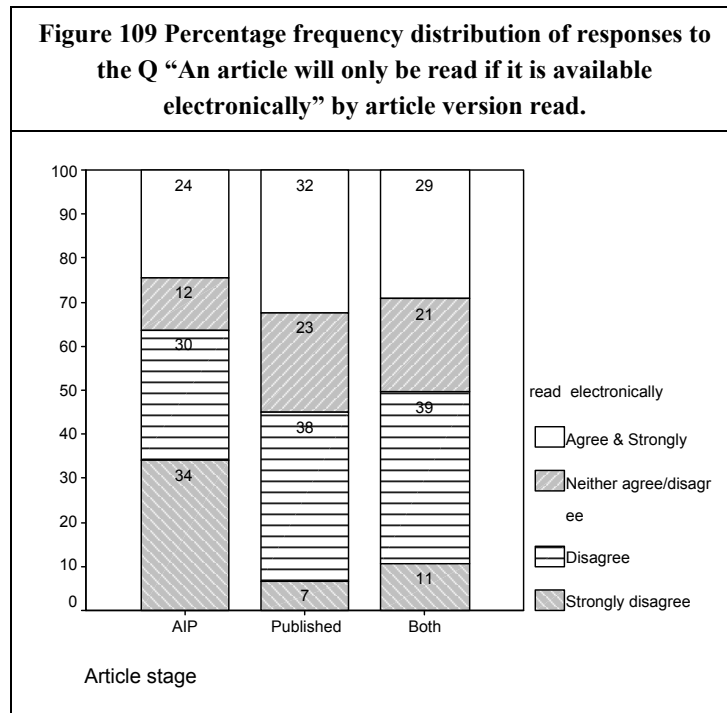
did so. Therefore people searching from work tended to view older material, maybe people do the more serious depth searching at work, which is logical.



4.3.1.2.2 Importance of the digital journal to the user

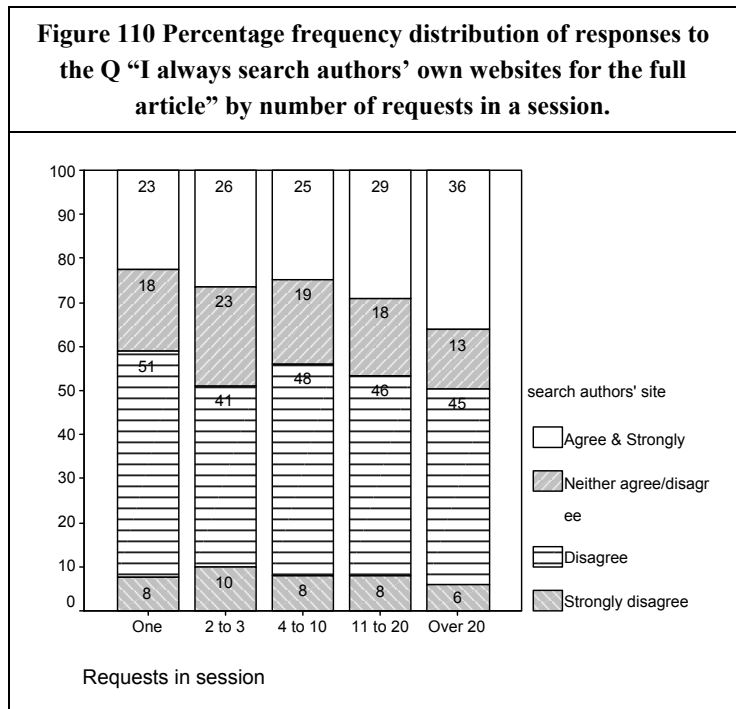
This analysis was based on responses to the statement “An article will only be read if it is available electronically” Answers were related to the use of articles in press, which is perhaps a currency indicator.

Figure 109 gives the percentage frequency distribution of responses to the question by what version of an article is read in a search session. Those just viewing articles in the in press stage in a session strongly disagreed with this statement. About a third (34%) strongly disagreed, compared to just 7% of those just viewing published material and 11% viewing both AIPs and Articles in a session. It seems that those users who appreciate the currency obtained from digital AIPs also appreciate the hard-copy. It could be that AIP users are more sophisticated users who would regard the statement as being simplistic.



4.3.1.2.3 Author’s websites as a source of articles

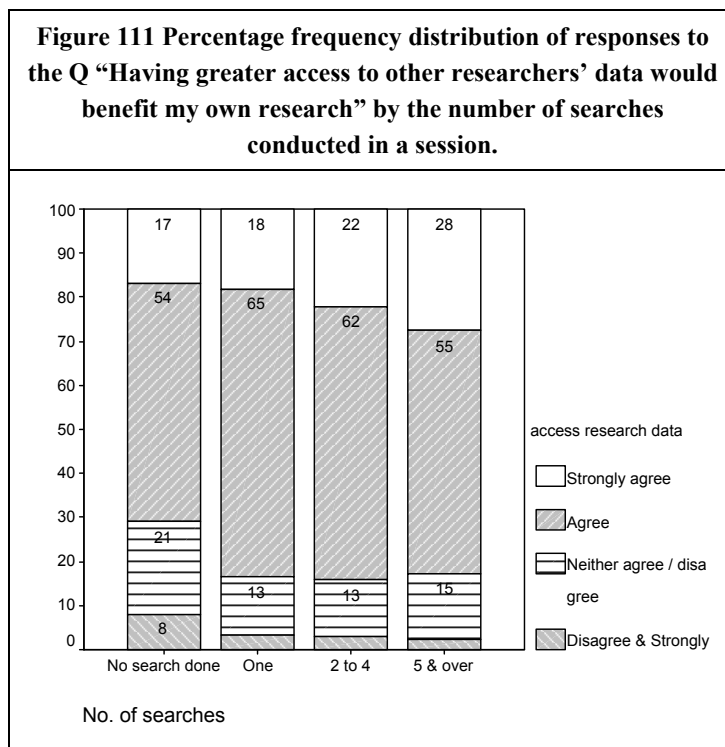
Figure 110 gives the percentage frequency distribution of responses to the question “I always search authors’ own websites for the full article” by the number of views in a session. Those respondents viewing more items in a session were more likely to strongly agree 36% did, as compared to 26% who agree but who just viewed 2 to 3 items in a session. It seems busy or active searchers of formal information sources, like ScienceDirect were more likely to seek out ‘free’ material from author’s websites.



4.3.1.2.4 Researcher’s raw data

Attitudes to greater access to other researchers’ data was investigated to see whether there were any differences according to the number of searches the respondent undertook in a session.

Figure 111 gives the percentage frequency distribution of responses to the question “Having greater access to other researchers’ data would benefit my own research” by the number of searches completed in a session (a busyness metric).



Those conducting more searches were more likely to strongly agree with the statement. About a quarter (28%) of those completing 5 or more searches strongly agreed with the statement compared to just 18% of those who just completed one search. More active users were more willing to explore other researcher’s datasets

The other side of the coin was also investigated and respondents were asked if they would make their data available and this was investigated to see whether there were any differences in attitude according the number of requests made, the number of searches made, and the number of journals they viewed in a session.

Figure 112 shows the percentage frequency distribution of responses to the question “I am willing to allow other researchers to access my raw research data” across the number of

Authors as users: a deep log analysis

requests made in a session. Those making more requests in a session were less likely to agree with the statement. About 37% of those making 20 or more requests in a session agreed, compared to 43% of those making 11 to 20 requests, 49% of those making 2 to 3 requests and 53% making 2 to 3 requests.

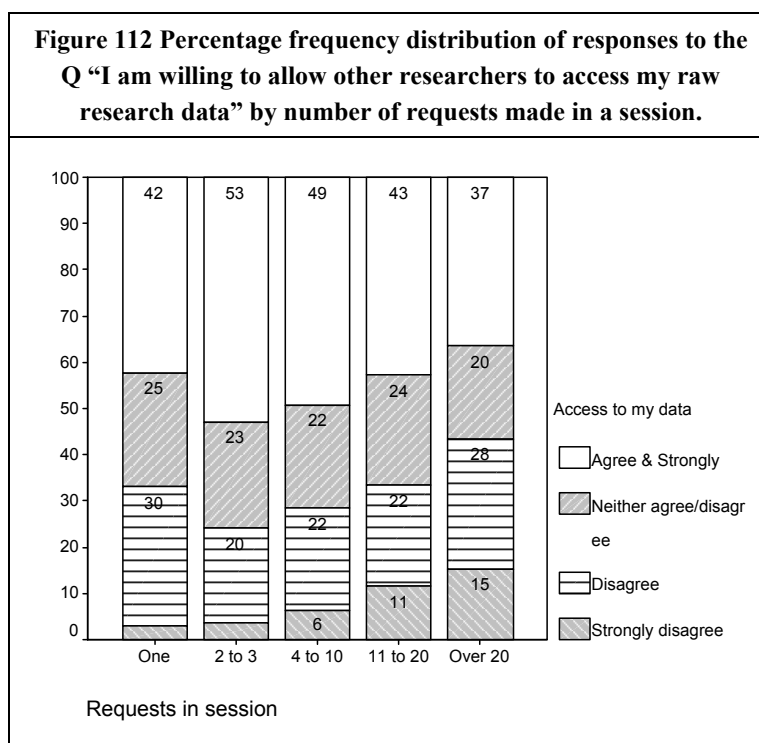


Figure 113 gives, for the same statement, a breakdown by the number of searches conducted in a session by the respondent. Those completing a greater number of searches were less willing to share their research data - 57% of those completing 5 or more searches said they would compared to 61% of those completing 2 to 4 searches and 66% who agreed of those just completing one search. The pattern is building that active scholars are less willing to share their data.

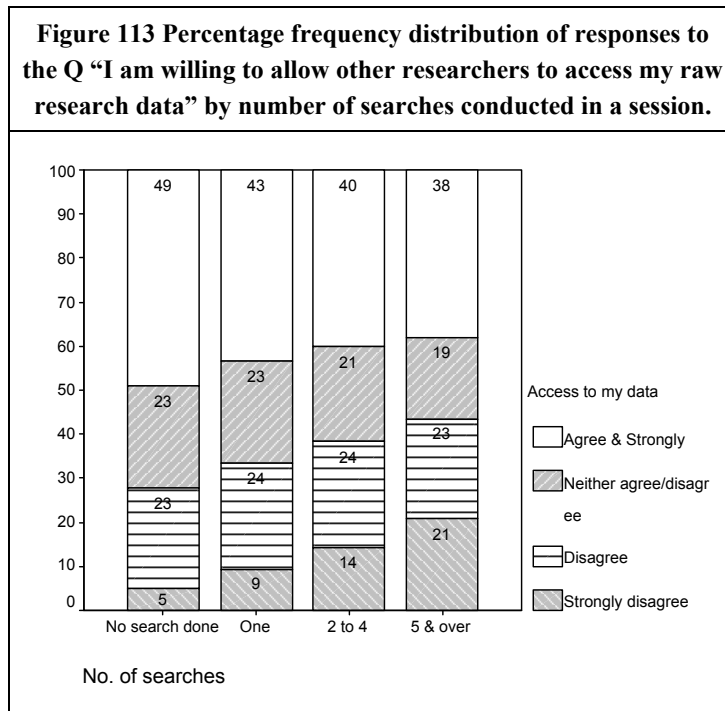
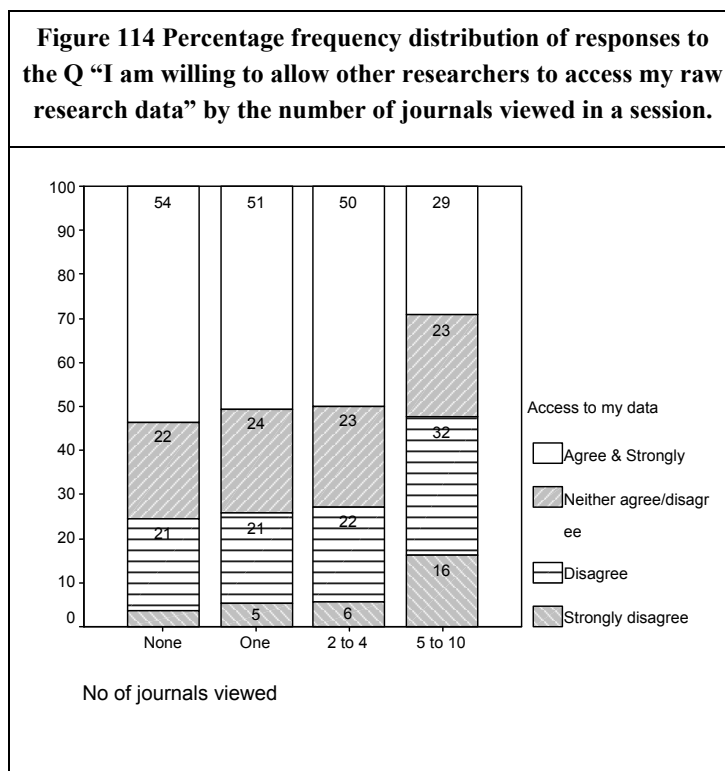


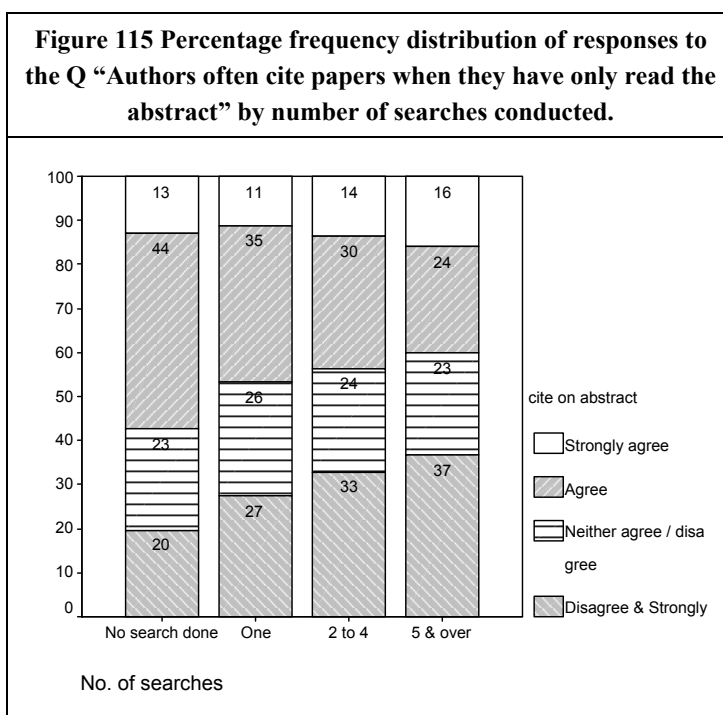
Figure 114 gives, for the same statement, a breakdown by the number of journals viewed by the respondent in a session. Those viewing more journals were less likely to agree to this statement. In particular those viewing 5 to 10 journals in a session were less likely to agree - half did as compared to three quarters viewing 4 or less journals.

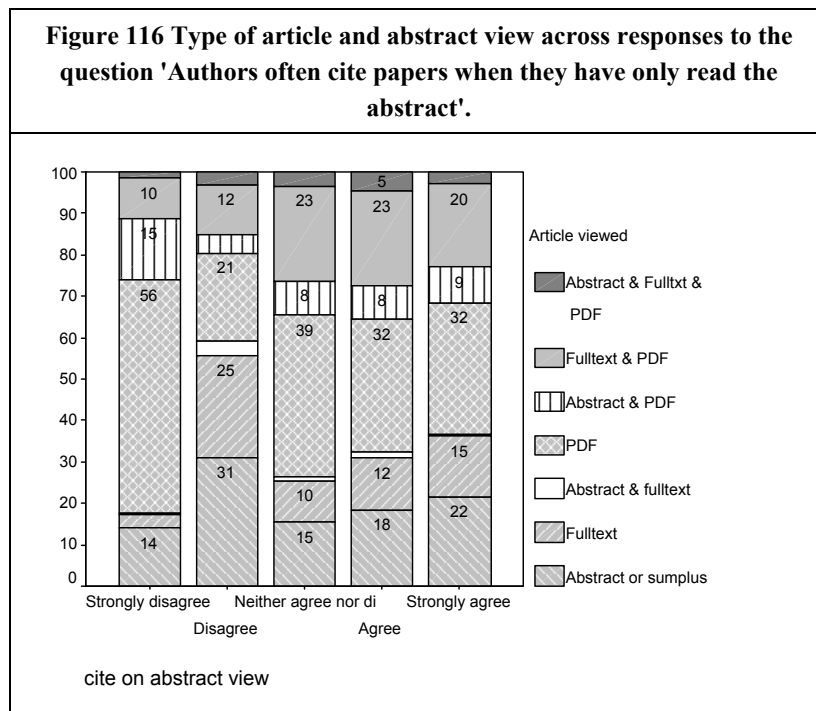


4.3.1.2.5 Citing behaviour

Figure 115 gives the percentage frequency distribution of responses to the question “Authors often cite papers when they have only read the abstract” by the number of searches completed in a session. Those respondents performing more searches were less likely to agree, the explanation possibly being that they were more experienced/serious authors.

Figure 116 looked at the type of article and abstract views by the response to the question. There proved to be little triangulation here with those who just disagreed with the statement being most likely, compared to other responses, to have a session where just an abstract was viewed about 31% did so. Just under a quarter (22%) of those who strongly agreed with the statement just had abstract sessions and this compares with 14% who did so who were in strongly disagreement with the statement.

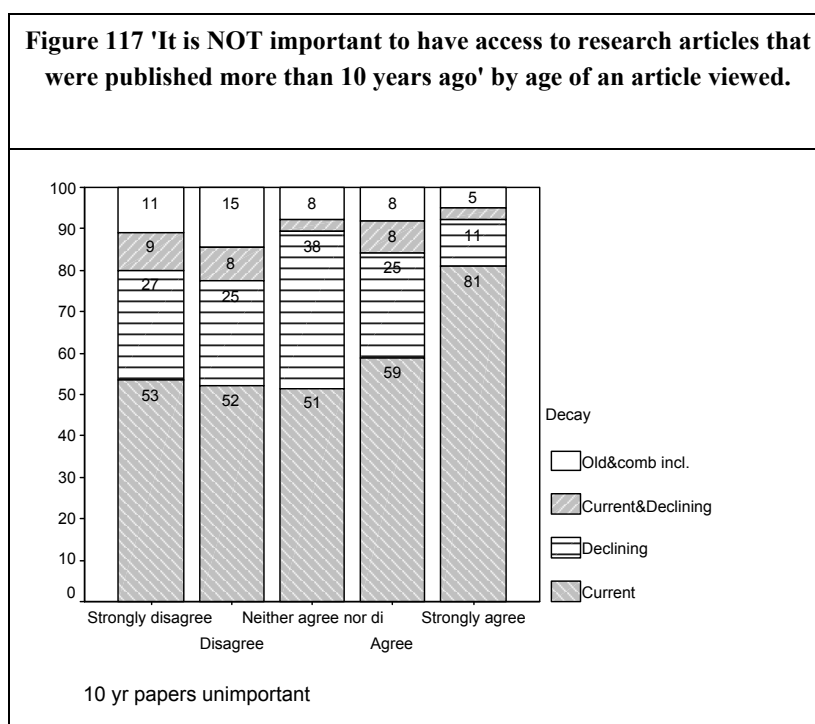




4.3.1.3 ARCHIVING

4.3.1.3.1 Article age

Figure 117 examines the relationship between responses to the question 'It is NOT important to have access to research articles that were published more than 10 years ago' to the age of the articles viewed. As might have been expected those who strongly agreed with the statement were most likely to engage in a session where just current material was viewed: 81% of sessions did so compared to about half of those sessions who strongly disagreed with the statement. Those who neither disagreed nor agreed were more likely, compared to other responses, to have a session where just declining material was viewed. Those strongly disagreeing or just disagreeing with the statement were more likely compared to those agreeing to have had a session where older material was viewed about 11 to 15% of these sessions included a view to older material compared to about 8% of sessions of those users agreeing with the statement.



4.3.1.4 FUNDING

4.3.1.4.1 Difficulty researching new topics

We sought to investigate whether those people who agreed with the statement "It is becoming increasingly difficult to carry out research in new and interesting areas", were more likely to complete more searches, more detailed searches, or broader searches (across subjects). In other words were these people ranging widely and thoroughly to find new material

4.3.2 Scholarly information seeking behaviour models

4.3.2.1 Online Behaviour – Model 1

This section seeks to identify core function online academic activity purely from looking at metrics generated from log files. It was decided to enter the various variables discussed into a factor analysis model to see if behavioural traits could be identified. Factor analysis identifies groupings of variables. Variables within the group are highly correlated while the resulting combinations are un-correlated and independent. The final (current) model is reported in Table 3 and included the viewing variables: decay rate of journal (current, declining and old), Journal Homepage and Issue page, Articles in print, Number of different journals viewed, abstracts viewed and full text views. The final selection of variables has inevitably played a part in the factors identified, as have the general emphasis of the dataset and the problems of extracting accurate metrics (caching and off site searching) from log files. Some variables, in this model,

could not be included together. For example views to PDF and HTML full text could not be included together. Here PDF views were left out.

Table 3: Factor analysis model 1 based on log metrics to identify academic search behaviour.

	Information collectors	Browsers	Updaters	HTML viewers
Initial Eigen values	3.0	1.7	1.27	0.9
Extraction %	29%	15%	10%	5%
Current	.486		.733	
Declining	.705		-.302	
Old	.451			
Journal Homepage	.292	.803		
Articles in print			.312	
Number Of Different Journals Viewed	.811			
Abstracts Viewed	.682			-.536
Full Text Views	.594			.309
Journal Issue		.641		

In all, four factors (user behaviours) were identified:

Information collectors: Users identified as having this type of behaviour looked at a number of different journals (.81), viewed declining material (.70) and, to a lesser extent, current (.49) and, old material (.45). They were likely to favour looking at abstracts (.68) and full texts (.59); though they also viewed PDFs.

Browsers: Users embarking on a Type 1 current awareness appeared to just update their knowledge by viewing journal homepages (.80) and Journal issues (.64). These users may of course come back for an information collection session at another time.

Updaters: Here users viewed current material (.73) were unlikely to view declining material (-.30) and looked at articles in print (.31).

HTML viewers – These viewers were not viewing abstracts (-.536) there was an increase tendency to view HTML full text (.31). However the fact that the PDF variable could not be included in this model it cannot be argued that these users did not view PDF items – they probably did. The model below attempted to correct for this.

These factors do not sum up all types of information collection behaviour. The four factor together account for just 60% of the variation that is 40% do some other than exhibit

Authors as users: a deep log analysis

information collection and current awareness behaviour. Further the fact that some variables could not be included in the model suggests a less than complete picture.

These factors do not sum up all types of information collection behaviour. The four factor together account for just 60% of the variation that is 40% do some other than exhibit information collection and current awareness behaviour.

4.3.2.2 Online Behaviour – Model 2

An additional factor analysis was undertaken, which just considered whether an event happened within a session or not, rather than just take into account the number of times that event happened overall. The final model is reported in Table 4 and included the viewing variables: decay period of journal (current and declining), Journal Issue page, Articles in print (AIP), whether user viewed a number of different journals, if user had used an abstract, had entered via a gateway or had completed a search. Again, the final selection of variables has inevitably played a part in the factors identified as have the general emphasis of the dataset and the problems of extracting accurate metrics (caching and off site searching) from log files. Again these issues probably say more about the accuracy of the split in extraction percentages, they are not particularly accurate, rather than the identified factors them selves.

Table 4: Factor analysis model 2 on log metrics to identify academic search behaviour.

	Gateway user	Searcher	Search & browser	Abstract viewer	Menu user
Eigen value	2.1	1.7	1.4	1.2	0.9
% variance	21.1%	13.5%	11.8	9.9	7.2
PDF/Full (0-non 1-either 2-both)	.725		.456	-.452	
Abstracts viewed				.690	.334
Via gateway	.714				
Search completed		.536	.364		-.393
Articles in print	-.365		.301		
Current material	.319	-.403	.618	.325	
Declining material		.713			.302
Use of different journals		.415	.318		
Journal issue	-.733		.330		.479

For this model five factors (user behaviours) were identified:

Gateway users: These users came in via a gateway (.71), did not use menus Journal issues (.73), on average (but not exclusively) preferred current material, would not generally view AIPs (-.37) and tended to view both PDF and Full text documents

Searchers: These users tended to use the search facility (.54), were average users of declining material (.71), below average users of current material (-.40) and tended to view a number of different journals (.42). These users would view items in either PDF or Full Text, but not both.

Search/browser users: these users tend to use a combination of the search engine (.36) and menus - journal issues (.33), they had a tendency to view current material - seemingly in both formats (PDF/Full .46) and they viewed a number of different journals.

Abstract viewers: These users predominately viewed abstracts (.69) and not PDFs (-.45) and tended to view current material (.33). Perhaps these users were non subscribers or came in via Google (bouncers?).

Browser: These users were gathering information just using menus -Journal issues (.5); perhaps, they have not been able to work out how to use the search facility because they seemed to be avoiding it (-.39) or they simply knew the parameters in which they wanted to conduct the search (i.e. a particular journal). They were viewing both abstracts (.33) and declining material (.30). These users viewed items in either PDF or Full Text but not both.

Both models give different views of the possible types of users and their online behaviour. The disadvantage of the first model is that not all the variables could be fitted in, and this gave a less than a full view; however, the full range of values of the included variables was included. The second model included more variables but at the expense of the range of values of the variables as here we were only interested if an event happened or not.

4.3.2.3 Online Behaviour - Model 1 informed by questionnaire results

It was decided to take the scores from Model 1 and run them through various user demographic and behavioural variables extracted from the questionnaire.

4.3.2.3.1 Factor 1: Information collectors

Users evidencing this kind of behaviour examined a number of different journals in a session, viewed declining material and, to a lesser extent, current and old material. They were also likely to look at a number of abstracts and favoured viewing in full text mode. These people ranged around widely for information.

Key characteristics:

Authors as users: a deep log analysis

- ✓ Users from Material Science and Mathematics were marginally more likely to search in this way;
- ✓ Hospital staff were marginally less likely to search and collect information in this way;
- ✓ Users from Germany, Netherlands and Canada were a little more likely to collect information in this way, while those from China were less likely;
- ✓ Were more likely to strongly agree with the statement that they “search authors' own websites for the full article”;
- ✓ Were less likely to search from home;
- ✓ Were on average more likely to agree that they “published to secure funding/tenure”.
- ✓ Were more likely to agree with the statement that “peer review does NOT improve article quality”.
- ✓ Were on average less likely to know about 'institutional repositories'

4.3.2.3.2 Factor 2 – Browsers

Users evidencing this form of information seeking behaviour appeared to obtain information by just viewing journal homepages and journal issues. These users may, of course, have come back at a later time for an information collection session.

Key characteristics

- ✓ Users from Business Management, Chemical Engineering, Chemistry, Environmental Science and Mathematics were identified as being on average more likely to engage in this form of behaviour, while Biochemistry, Biological Science and Material science users were less likely to.
- ✓ Older users, particularly those aged over 65, were less likely to update their knowledge in this way. Perhaps they had reached the end of their careers and thus did not need to seriously keep up to date?
- ✓ Women were less likely to exhibit this form of behaviour;
- ✓ Post graduates and senior researchers were marginally less likely to exhibit this form of behaviour.
- ✓ Users from China, Germany and Spain were marginally more likely to behave in this way, while respondents from the Netherlands, UK and US were less likely to.
- ✓ Were less likely to agree with the statement articles will “only be read electronically”.
- ✓ Were more likely to agree with the statement “conferences, bulletin boards are NOT important in scholarly publishing” - that is this group on average were less likely to use them.

- ✓ Were on average less likely to agree with the statement “The publisher adds little value”.
- ✓ Were on average more likely to agree that they “published to secure funding/tenure”.
- ✓ Were more likely to agree with the statement that “peer review does NOT improve article quality “.

4.3.2.3.3 Factor 3 - Updaters

In the case of updating behaviour users were viewing current material, were unlikely to view declining material and would view articles in print. These users viewed full text content in PDF.

Key characteristics

- ✓ Chemical Engineering and Mathematics users were marginally less likely to exhibit this form of behaviour;
- ✓ Older respondents (65 & over) were marginally more likely to behave in this manner;
- ✓ Women tended to marginally favour this form of behaviour;
- ✓ Post graduates tended to behave in this way;
- ✓ Users from the Netherlands and Spain marginally tended towards this form of behaviour, while users from Germany, Canada and France were marginally less inclined to behave in this way.
- ✓ Those behaving in this manner were more likely to agree with the statement that they “search authors' own websites for the full article”,
- ✓ Were more likely to agree with the statement Quality of an article is determined by the journal
- ✓ were on average more likely to agree that they published to secure funding/Tenure
- ✓ Seemed less likely to agree with the statement that “peer review does NOT improve article quality”.
- ✓ Were on average more likely to know about 'institutional repositories'

4.3.2.4 Online Behaviour - Model 2 informed by questionnaire results

Key characteristics associated with this form of behaviour were as follows:

In general users were **less** likely to agree with the statement “Quality of an article is determined by the journal”

Gateway users

- ✓ Seemed less likely to agree with the statement that “peer review does NOT improve article quality”

Authors as users: a deep log analysis

- ✓ Were more likely to agree that the publisher adds little value to the article. An explanation might lie in that these users were coming in via a gateway and they were searching the gateway site for content from the publishers' site. These users may not associate the finding of content, or indeed the authority of content, with the publisher.
- ✓ Were on average more likely to agree that they published to secure funding/tenure.
- ✓ Were on average more likely to agree that authors will pay to have their articles published

Abstract viewers

- ✓ Were marginally less likely to agree that the publisher adds little value to the article.
- ✓ Were on average less likely to agree that they published to secure funding/tenure.
- ✓ Were on average less likely to agree that authors will pay to have their article published
- ✓ Seemed less likely to agree with the statement that they were "unable to review the literature as thoroughly due to time constraints".
- ✓ Were more likely to visit just once and were less likely to be regular users.
- ✓ Searchers
- ✓ Were on average less likely to agree that they published to secure funding/tenure
- ✓ Were on average less likely to agree that authors will pay to have their articles published
- ✓ Seemed less likely to agree with the statement that they were unable to review the literature as thoroughly due to time constraints
- ✓ Were on average less likely to know about 'institutional repositories'.

Searcher/browser

- ✓ Were less likely to visit just once and were more likely to be regular users.

4.3.3 Returnees

Returnees were not one of the models identified as part of the factor analysis but we felt, because of the loyalty and satisfaction they have demonstrated, they should be treated as a special group, and we sought through questionnaire responses to understand those people visiting more often.

Characteristics:

Journal title was a key a factor. In regard to the question "the quality of an article is determined by the journal", users visiting more regularly (6 or more times) were more likely to agree with this statement and less likely to disagree with it.

Generally peer review supporters. Those returning more frequently to the site were less likely to agree to the statement “Peer review does NOT improve article quality”.

Believe in the value of digital journals. In regard to the question “articles will only be read electronically” those visiting more regularly, 6 or more times, were more likely to agree with the statement and less likely to disagree with it.

Not interested in the author’s website. With regard to the question do you “search authors' own websites for the full article”, users visiting more regularly were more likely to disagree with this statement – about half (52%) of those visiting 15 or more times disagreed compared to 28% of those just visiting once.

Split regarding user input. With regard to the statement “published articles can be revised in light of comments posted online by readers (e.g. continuous review)”, views tended to become more polarised as the number of times a user returns to the site increases – both those in agreement and disagreement increased.

5. General Conclusions

This was an exploratory and innovative study which sought to go where no other research project had gone before, that is link usage and search logs with attitudinal and demographic data obtained from questionnaire to obtain a fuller understanding of the virtual scholar. The methodology clearly created a rich picture of the users of ScienceDirect, albeit a relatively small group of them in deep log analysis terms, but certainly not in general social survey terms, where the sample size is quite impressive. The characteristics of the sample population is significant and is worth repeating here: the highest number were between the ages of 36 and 45; men out numbered women 3 to 1; Medicine and the Social Sciences were well represented; about three quarters of respondents were academics; senior professors/heads of department accounted for 43% of users, a fact that probably reflects the fact that this is the group that are most active in producing academic papers; the US accounted for nearly a third of the sample. However, this distribution is in fact representative of the scientific scholarly community at large.

The main findings of the study have been described in the Executive Summary and just the most significant findings will be mentioned here. The key findings in regard to usage and searching were:

- *The very significant levels of author/user diversity discovered.* There were very real differences between various types of user, especially in regard to their subject field; academic status and geographical location, but also sometimes in regard to age, gender, and organisation affiliation This highlights the great danger of trying to generalise data on the back of hundreds of thousands of users (lumping Nobel prize

winners and undergraduates together) and points to the need to always view usage data in the context of user data. Clearly in terms of digital library or platform design one size does not fit all.

- *The high rate of loyalty/dependence shown by users.* The number of users returning was greater than we have seen in studies of digital journal libraries elsewhere and probably points to the fact that: a) scientists are generally more frequent visitors because of the nature of their subject field; b) ScienceDirect has a more loyal body of users, due to the titles it has, the kind of people that use it and the quality of the product. Returnees clearly represent satisfied and loyal users and we have looked at the qualities and attitudes of this particular user group and they appeared to be more conservative in their attitudes towards the scholarly communication system.
- *The depth to which the resource was used.* Just over a third of ScienceDirect users viewed 4 to 10 items in a session and 15% viewed more than 10 items. By comparison with other journal databases this represents generally high levels of penetration.
- *Length of an article.* People spent more time reading shorter articles online than long ones and the longer the article the more likely it was only going to be read in abstract form.

The most important finding regarding scholarly behavioural and attitudinal traits concerned the identification of a number of important information seeking models derived from factor analysis. The key behavioural models follows with an attribution of user demographic characteristics associated with that model:

Information collectors: Users evidencing this kind of behaviour examined a number of different journals in a session, viewed declining material and, to a lesser extent, current and old material. They were also likely to look at a number of abstracts and favoured viewing in full text mode. These people ranged around widely for information. Those adopting this form of behaviour were likely to come from Material Science and Mathematics, live in Germany, Netherlands and Canada, and work in hospitals,

Browsers. Users evidencing this form of information seeking behaviour obtained information by just viewing journal homepages and journal issues. Those adopting this form of behaviour were likely to come from Business Management, Chemical Engineering, Chemistry, Environmental Science and Mathematics, live in China, Germany and Spain and were older.

Updaters. In the case of updating behaviour users viewed current material, were unlikely to view declining material and viewed articles in print. These users viewed full text

content in PDF. Those adopting this behaviour were likely to come from the Netherlands and Spain.

Other behavioural groupings were identified using another factor analysis and a description of these groups follows together with an attitudinal analysis of users belonging to some of these groups:

Gateway users. These users came in via a gateway, did not use menus (journal issues), on average preferred current material, would not generally view articles in press and tended to view both PDF and Full text documents. Their key attitudinal characteristics were:

- ✓ Less likely to agree with the statement that “peer review does NOT improve article quality”
- ✓ More likely to agree that the “publisher adds little value to the article”.
- ✓ More likely to agree that they published to secure funding/tenure.
- ✓ More likely to agree that “authors will pay to have their articles published”

Searchers. These users tended to use the search facility, were average users of declining material, below average users of current material and tended to view a number of different journals. These users viewed items in either PDF or Full Text, but not both. Their key attitudinal characteristics were:

- ✓ Less likely to agree that they published to secure funding/tenure
- ✓ Less likely to agree that “authors will pay to have their articles published”
- ✓ Less likely to agree with the statement that they were “unable to review the literature as thoroughly due to time constraints”
- ✓ Less likely to know about 'institutional repositories'.

Search/browser users. These users tended to use a combination of the search engine and menus - journal issues, they had a tendency to view current material - seemingly in both formats (PDF/HTML) and they viewed a number of different journals. These users were less likely to visit just once and were more likely to be regular users.

Abstract viewers. These users predominately viewed abstracts and not PDFs and tended to view current material. Their key attitudinal characteristics were:

- ✓ Less likely to agree that the “publisher adds little value to the article”.
- ✓ Less likely to agree that they “published to secure funding/tenure”.
- ✓ Less likely to agree that “authors will pay to have their article published”
- ✓ Less likely to agree with the statement that they were “unable to review the literature as thoroughly due to time constraints”.

- ✓ More likely to visit just once and were less likely to be regular users.

Browsers. These users gathered information just using menus – journal issues avoided using the search facility. They viewed both abstracts and declining material. These users viewed items in either PDF or Full Text but not both.

The combination of logs and questionnaire data has provided all kinds of interesting results regarding authors and their scholarly practices and views. It has also raised questions that need to be asked in the next Author questionnaire. This was a pilot study which has demonstrated profitable and unprofitable lines of investigation and raised questions that urgently need answering. With this knowledge and a methodology in place now for investigations of this power a similar study on a bigger sample population, say, 3000 authors. ScienceDirect has a huge lead over its competition in understanding the user and this would ensure it maintained its lead.

6. References

Fieber, J. (1998), *Browser Caching and Web Log Analysis*, Available at: <http://ella.slis.indiana.edu/~jfieber/papers/bcwa/bcwa.html> (visited 9 September 1999).

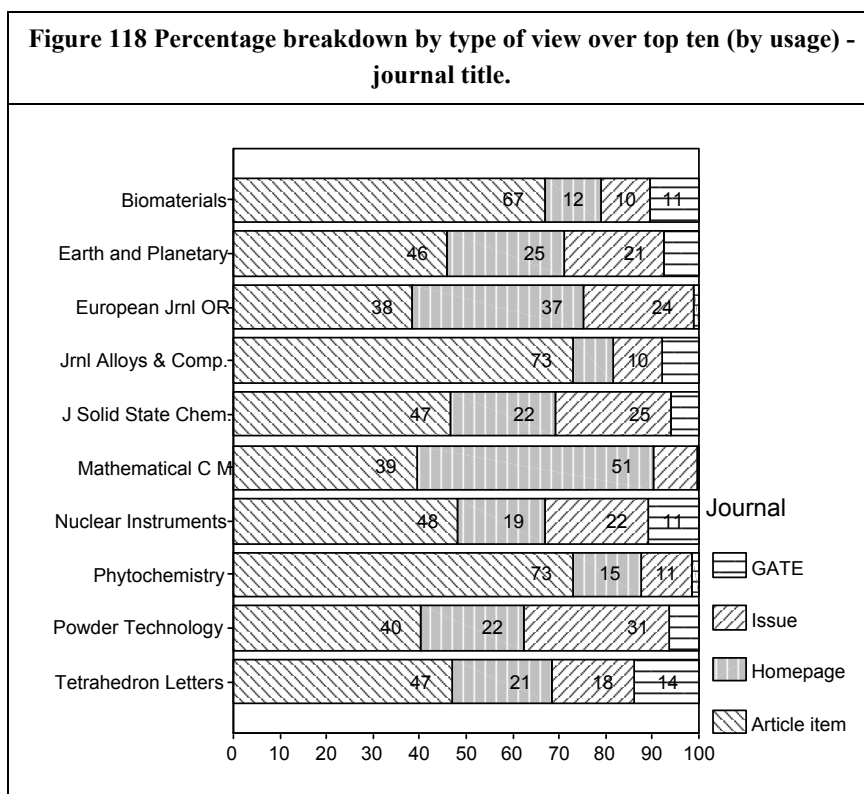
Mabe, M.A. & Amin, M. (2002), “Dr Jekyll and Dr Hyde: Author Reader Asymmetries in Scholarly Publishing”, *Aslib Proceedings*, 54(3), 149-175.

Tauscher, L., & Greenberg, S. (1997), “How people revisit web pages: empirical findings and implications for the design of history systems”, *International Journal of Human-Computer Studies*, 47,97-137.

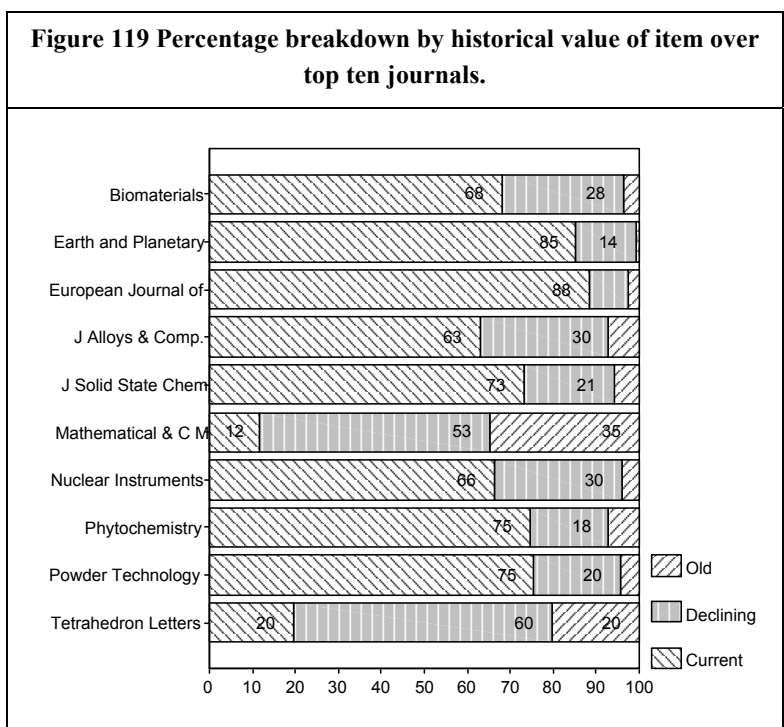
7. Appendix 1: Supplementary analyses

7.1. Top Ten Journal titles

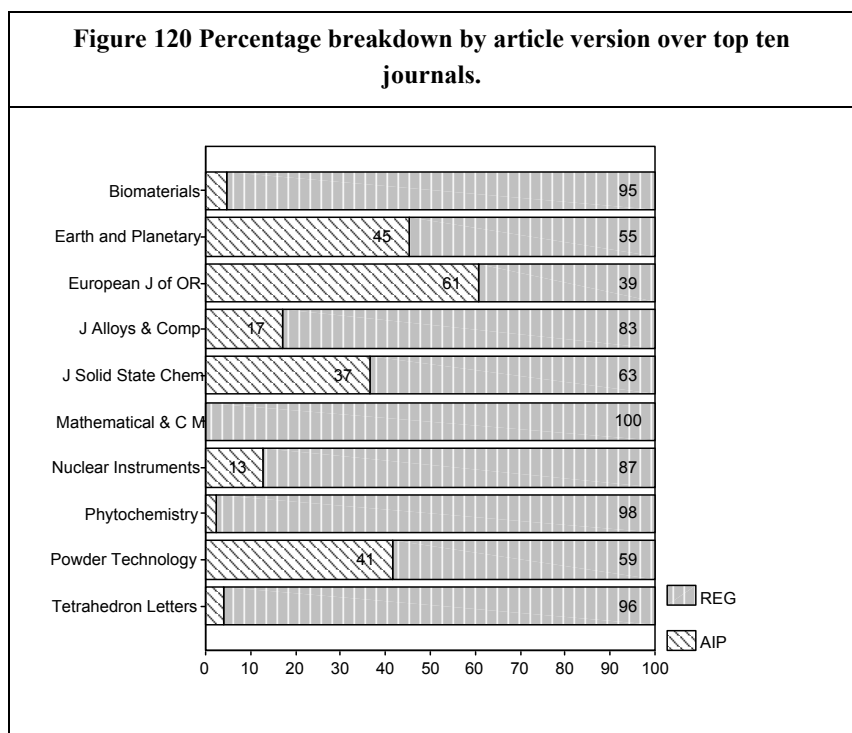
With regard to the top ten journal titles and type of page view (Figure 118) Powder Technology, Journal of Solid State Chemistry and European Journal of Operational Research recorded the highest views to issue pages (respectively, 31%, 26%, 24%), relatively high views to the journal homepage (22%, 22%, 37%) and relatively low views to article items (40%, 47%, 36%). Mathematics & Computer Modelling recorded particularly high views to the homepage (51%). Biomaterials (67%), Journal of Alloys and Comp. (73%) and Phytochemistry (73%) had high views to article items.



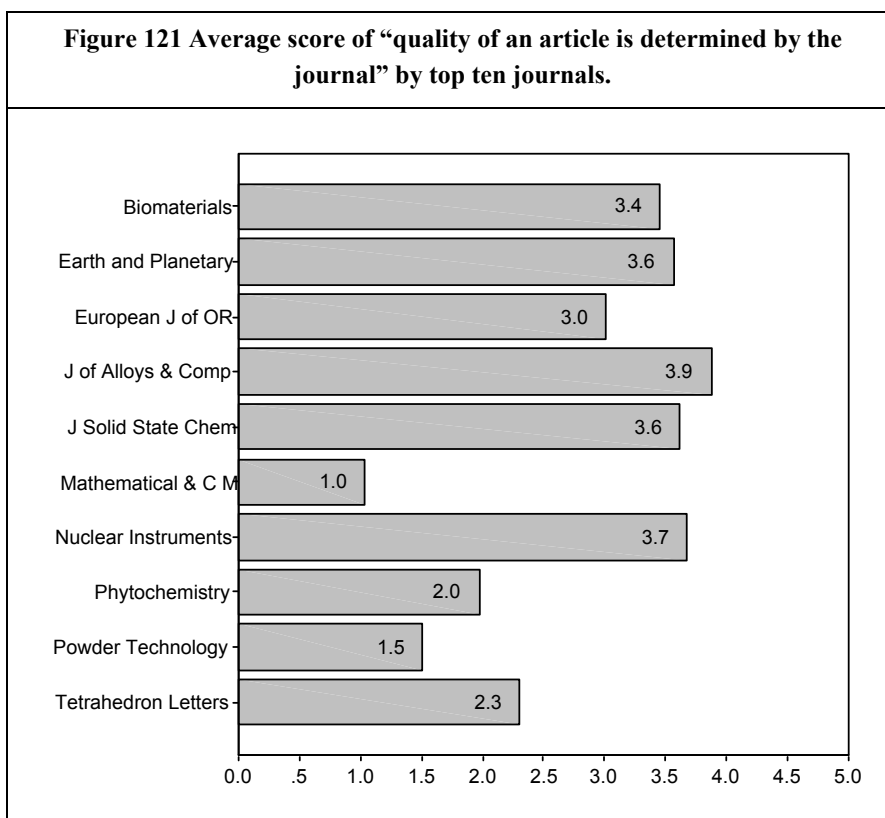
In terms of individual journals and views to historical items (Figure 119) users of Mathematics and Computer Modelling and Tetrahedron Letters made extensive use of historical items and between 80 to 90% of views were to declining (53%, 60%) or old (35%, 20%) material. Almost all the views to the European Journal of Operational research (88%) and Earth and Planetary (85%) were to current material.



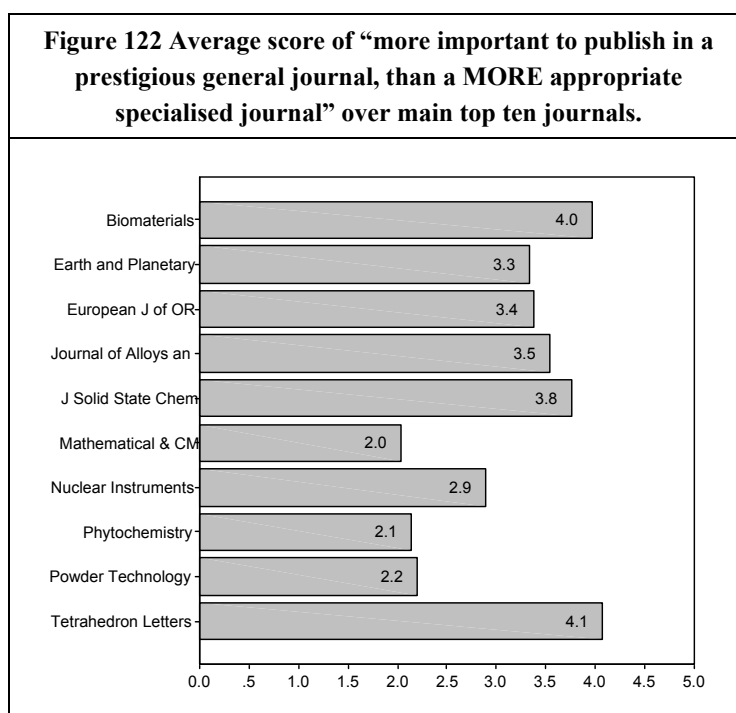
In terms of accesses to in print articles (Figure 120) the European journal of Operational Research (61%), Earth and Planetary (45%) and Powder Technology (41%) scored highly on this metric, while Biomaterials (5%), Mathematics and Computer Modelling (0%) and Powder Tetrahedron Letters (5%) scored lowly.



Combining questionnaire data with usage data, specifically the responses to the question Q1 the “quality of an article is determined by the journal”, those authors using the journals Journal of Alloys & Comp. (3.9), Nuclear Instruments (3.7), Earth and Planetary (3.6) and Journal of Solid State Chemical were likely to agree while those using Mathematics and Computer Modelling (1.0), Powder Technology (1.5) and Photochemistry (2.0) were likely to disagree. (Figure 121)



While for Q8, the “more important to publish in a prestigious general journal, than a MORE appropriate specialised journal” (Figure 122), those viewing Biomaterials (4.0), Tetrahedron Letters (4.1) and Journal of Solid State Chemistry (3.8) were likely to agree, while those accessing Mathematics and computer modelling (2.0), Photochemistry (2.1) and Powder Technology (2.2) were less likely to agree.

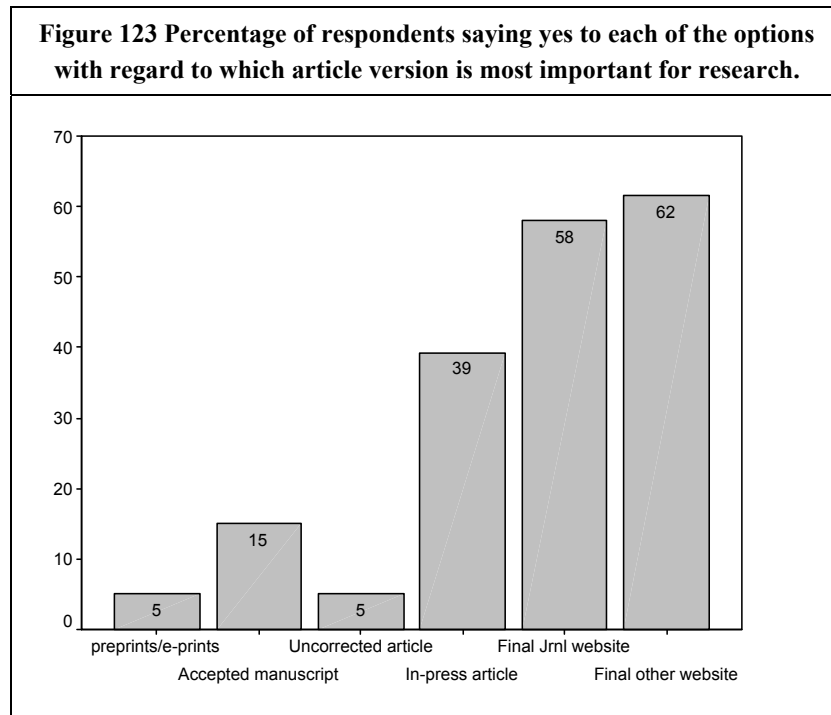


7.2. Article versions analysis

Authors were asked about their views concerning the digital existence of different versions of the same article – a possible digital nightmare for the user. A journal article can undergo many stages and be available to researchers at different points in the chain and in differing locations. Respondents were asked which version of an article they considered the most important for research. Unfortunately the questionnaire did not give a range over which respondents could express their importance ranking for each option but instead were asked to tick, a single box only, from a list of alternative options which they considered the most important for your research. The options were: non peer-reviewed pre-publication drafts (often called preprints or e-prints); author’s final manuscript version after acceptance; uncorrected proof of the article (after acceptance - normally a PDF and typeset by the publisher); in-press article (Author’s corrected proof version available via the journal’s web site in an “in press” location); final published article (available via the journal’s website including volume, issue and page numbers) and final published article (generally a PDF; available via a website other than the journal’s). Respondents could choose more than one form. Figure 123 provides the results. Respondents not ticking yes to any of the versions were coded as missing and were not included in the total upon which the percentages were calculated.

Most respondents (62%) said that for research purposes, they preferred to view the final published article (generally a PDF; available via a website other than the journal’s). Over half (58%) thought the final version on the publishers web site (note respondents could select more

than one option) was what they wanted, 39% said the article in press, 5% an uncorrected piece, 15% the authors accepted manuscript and 5% the non peer-reviewed pre-publication draft.

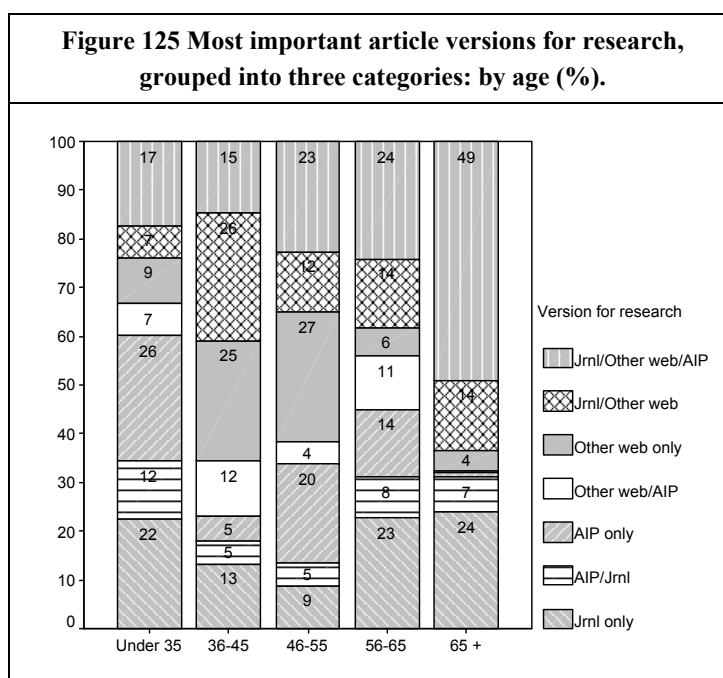
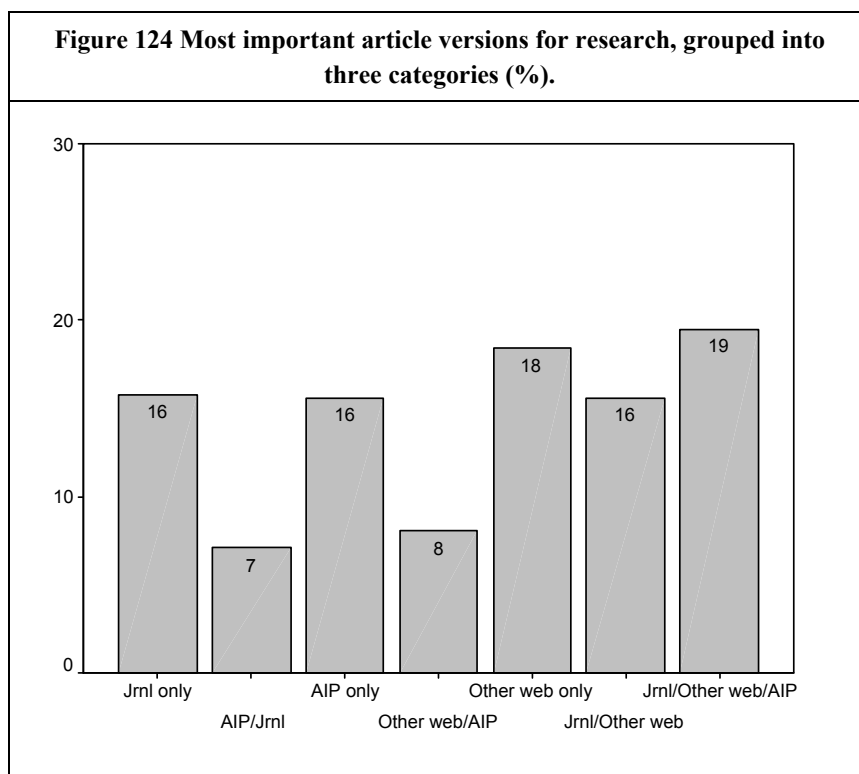


Authors could select more than one version and might prefer one or a combination of versions when conducting research. To examine this in more detail it was decided to group the options into three groups. The first group, labelled articles in press (AIP), grouped pre-prints, accepted material, uncorrected articles and in-press articles together. The second and third groups were the options final published version via the journal’s website (Jnl) and final published viewed on a web site other than the publishers (Other web). Figure 124 gives the percentage distribution. About 1 in 5 (19%) authors used all three article groups for research purposes. Approximately 18% said that only other web final versions were the most important for research purposes; while 16% equally thought that the publisher’s material version and AIP editions were the most important. Seven percent rated the combination of the publisher’s material version and AIP editions and 8% rated other web site final version and AIP editions as important.

In terms of age (Figure 125), older users were more likely to rate a combination of all three formats for research; 49% of those aged 65 and over said so compared to about 17% of those aged 35 and under. Respondents aged between 36 and 55 were more likely to rely on just using other web sites for published articles - about a quarter did so compared to less than 10% for other age groups. Those 35 and under appeared more likely to use AIP versions, and about

Authors as users: a deep log analysis

a quarter (26%) did so. Use of the journal's web site versions appeared to decline with age, except for those aged over 55.



In terms of occupational status (Figure 126), comparing heads of Departments, researchers and senior researchers only, about a third of senior researchers just used other web sites to access articles, compared to 9% of researchers and 16% of Heads. Perhaps, senior researchers had less time? Senior researchers were least likely to rate as important a combination of all three article stages; 12% said so compared to about a quarter of heads of department (23%) and researchers (24%). Research staff were most likely to rate AIP versions 25%, did so compared to just 6% of senior researchers and 19% of department heads.

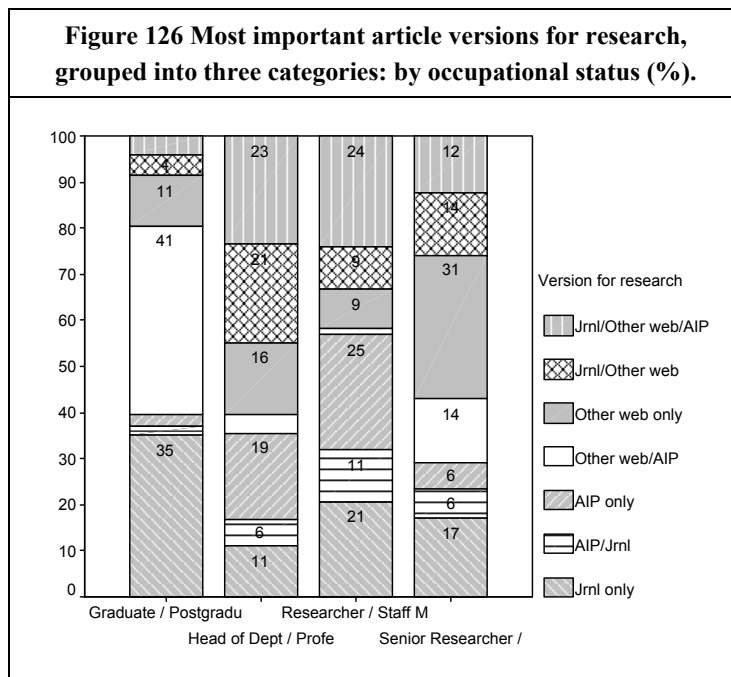
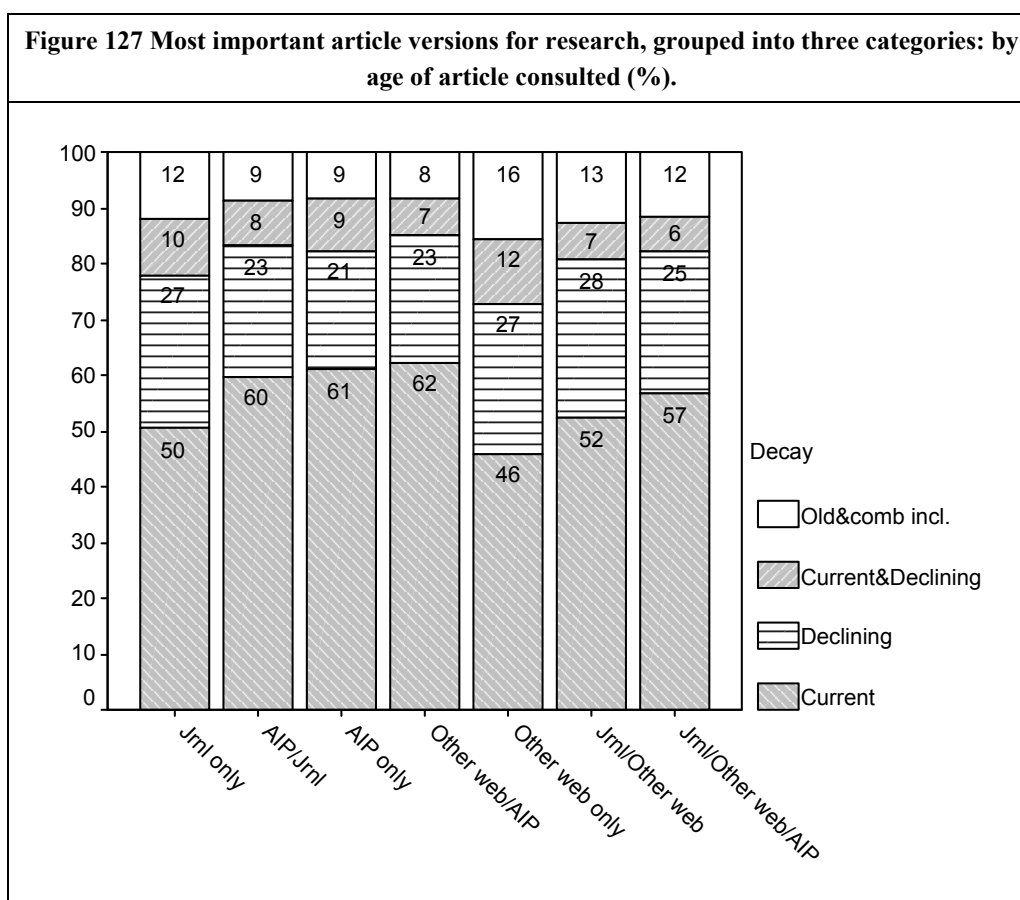


Figure 127 gives the distribution percentage share of age of article by which published version (grouped) was preferred for research purposes. There is a tendency for those respondents rating other web sources and the journal's final version to be more likely to look at older material.



Analysis examining the preference for article version and users depositing of their own material.

It was decided to compare, contrast and look for associations (using principal component analysis) between responses to the question regarding which versions of articles they considered the most important for research and how respondents deposited their own material. In part the questions are two sides of the same coin, one asks what versions they used and the other what type of version they deposited. Respondents had five options with regard to depositing their own material these were: deposited an article in an institutional repository; deposited an article in a subject repository (i.e. Pubmedcentral, ArXiv); posted an early version of article on personal website; posted the final published version on personal website; published it in an open access journal so that it was freely available on the world wide web. It can be seen that authors and readers – and of course some times they are the same thing, are faced with a complex and possibly confusing digital world. The principal component analysis explained just less than three quarters (71%) of the variance. However, it should be remembered that users in each case could respond to more than one option and to each option the user had either the option to tick or not to tick. The interpretation of principal component

analysis is limited to the variables included and this might result in a false or distorted picture if significant other variables have been left out. Hence the results should be considered in the light of other research material. The results are reported in Table 5.

Table 5: Principal component analysis rated importance of article type and how respondents deposited their own material.

	Component					
	1	2	3	4	5	6
71%	16%	15%	13%	10%	10%	8%
preprints/e-prints	.34	.33			.61	
Author's accepted manuscript	.51		.47			.46
Uncorrected article proof	.71		.31			
In-press article		.48	.54	-.40		
Final published article via the journal		-.35	.53	.52		
Final published article via other website	.43	-.35		-.30	.35	
Deposited in an institutional repository		.55	-.30	.34	.45	
Deposited in a subject repository		.71				
Posted an early version on your website	.49			.52		-.44
Posted final published version on your website	.55	.34				-.36
Published in open access journal	.31		-.55	.29	-.37	.39

For the first component: this factor represents a “**do as I do**” attitude, as those that rated uncorrected article proofs (.71) and author accepted manuscripts (.51) as important for research tended to post an early version of their article on their website (.49), posted final published version on their website (.55) and to a lesser degree published in open access journal (.31). That is they tend to rate as important what they did with their own material. These users also rated the final version on another website as important for research (.43) and preprints (.34). These users were not particularly journal committed, either in the case of open access or print journals. They probably saw the availability and uploading of digital article resources as a relatively costless procedure.

The second group is difficult to characterize, but appears to be **subject repository committed**. These users rate highly the depositing of their material in a subject (.71) and

institutional (.55) repository, they do not really rate for research purposes either the Journal final version (-.35) or final version from another website (-.35) but instead rate in-press articles (.48) and pre prints (.33) they appear marginally more likely to post a final published version of their material on their own website (.34). These users do not appear journal committed and probably believe that structured journals will disappear and be replaced by repositories. Doubts are clearly there though, with the relatively high score for in-press versions compared to pre print ones.

The third group seems to be formed of the **Official version committed ones**. These people will access the final version from the publishers web site (.53) but will also access in-press articles (.54), the authors accepted manuscript (.47) and uncorrected proof (.31). They don't tend to deposit material in an institutional repository (-.30) and will not publish in open access (-.55). These users are journal committed.

The fourth group appear to “**play it all ways but prefer a publisher version in the end**”. For research they rate the published version (.52); however, they'll post an early version on their website (.52) and are willing to publish in an open access journal (.29). They don't rate the in-press version (-.40) or final versions from other websites (-.3); however, they will deposit in institutional repository (.34). Their research might not be literature led (they are not interested in early versions), but they are interested in publishing in peer reviewed journals and need the references of published material to cite and be cited. These users are journal committed because it suits them and it is how they see the system working – through citations etc.

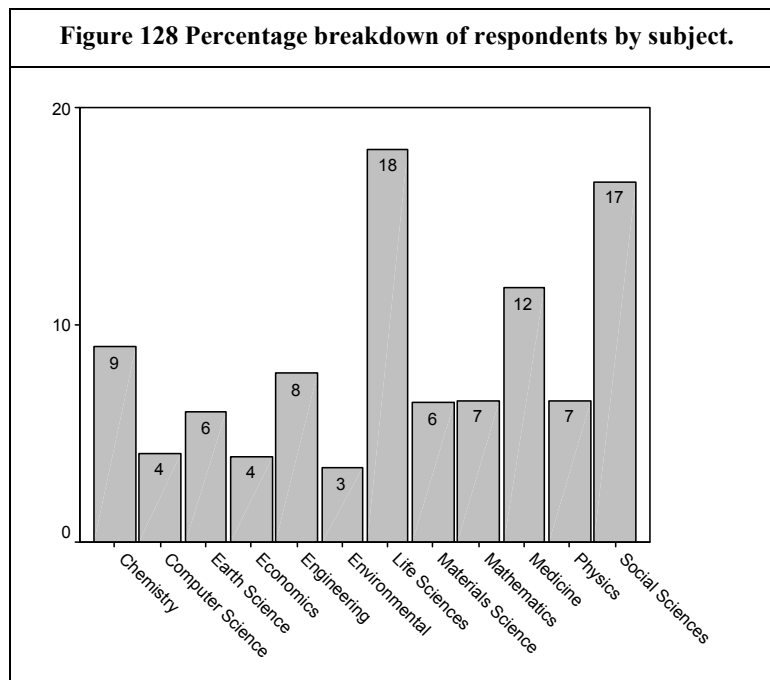
The fifth group are “**early version users**”. For research this group rate pre-prints highly (.61) however they have deposited material in an institutional repository (.45) but tend not to publish in open access (-.37) and will use other web sites to view final versions (.35). These users believe in their own individually honed research methods, they “know” their websites and they feel this gives them an edge in keeping up to date. Yes, journal committed.

The last group is difficult to characterize however, but they seem to be **committed to open access journals**. For research purposes they rate the author's accepted manuscript (.46) and they are likely to have published in an open access journal (.39), but are less likely to make available versions either early (-.44) or final (-.36) on their website. These users are committed to a formal journal structure albeit an open access one.

8. Appendix 2: Characteristics of sample

In all there were 757 authors who completed questionnaires which could be matched to the ScienceDirect usage logs. These users viewed 110,000 items.

In terms of age, the highest proportion (33%) of respondents were between 36 and 45 years old and the age groups under 36 (24%), 46-55 (30%) and 56-65 (12%) were also well represented. However, less than 10 users were under 26, not surprising given that the questionnaires went to authors of academic papers. In regard to gender there were almost three times as many men in the sample. Figure 128 shows that the most popular discipline to which users belonged was Life Sciences (18%), and this was followed by the Social Sciences (16%)



In regard to type of organisation to which they belonged, most users came from universities and colleges and accounted for over three quarters (78%) of all respondents. Research institutes were also well represented (17% of respondents).

In terms of the academic status of respondents most of the respondents were either professors (42%) or researchers of some sort - researchers (23%) and senior researchers (32%). The number of students (graduates/postgraduate) was relatively low at 53, and made up just 3% of respondents.

The geographical spread of the sample was quite wide, although respondents from the North America (33%) and Western Europe (38%) were clearly in the majority. About 17% came from Asia and 6% from Eastern Europe (Figure 129). This generally reflects the global distribution of papers (ISI) and hence is generally representative.

