

**Improving the comparability of usage statistics and
the implementation of unique article identifiers
(*Publisher Metadata and Interoperability Projects II*)**

Final Report

**Peter T Shepherd
Project Director
COUNTER**

May 2006

Table of Contents

	Page
1. Project Team and Acknowledgements.....	3
2. Executive Summary.....	4
3. Background.....	5
4. Aims and Objectives.....	5
5. Methodology.....	6
6. Implementation.....	7
7. Outputs and Results.....	16
8. Outcomes.....	17
9. Implications.....	17
10. Recommendations.....	18
11. References.....	18
12. Appendices.....	19

1. Project Team and Acknowledgements

a. Project team

The Project Team were: Peter Shepherd (COUNTER, Project Manager): Marthyn Borghuis (Elsevier): Oliver Pesch (EBSCO): Angela Conyers and Pete Dalton (Evidence Base).

b. Acknowledgements

This project has been funded by JISC under the auspices of the Publisher Metadata and Interoperability Projects II programme and we are most grateful for this support. In particular we would like to express our thanks to Christine Baldwin for her guidance on the project and to Phil Davis (Cornell University) for his input into the project. It is also important to acknowledge the resources devoted to this project by Elsevier and EBSCO, as well as the work of Publishers A to E in running the tests of the unique article filter.

2. Executive Summary

The aim of the COUNTER Filter project was to develop a data filter that would deal with the possible inflation of usage statistics due to interface effects, assess its implications for the application of unique article identifiers, such as DOIs, and to propose how the filter could be implemented by vendors. This overall aim translated into three specific objectives:

- a. The development of a filter to be applied to usage data that would dampen or compensate for the inflationary effects of certain vendor interface configurations.
- b. An assessment of current vendor practice regarding implementation of unique article identifiers, such as DOIs
- c. Propose how the filter could be implemented by vendors and recommend appropriate modifications to the COUNTER Code of Practice to the COUNTER Executive Committee.

In the early stages of the project it was decided to test two data filters, as no experimental evidence existed to indicate if either approach would work in practice. The two filters tested were:

- the 'unwanted html' filter, designed to filter out requests for full-text html articles deemed not to have been specifically requested by the user.
- the 'unique article filter' (UAF), designed to provide a count for the number of requests for unique articles (irrespective of format) in a given session.

The project was divided into three Phases and the overall approach taken had two main strands, which were carried forward in parallel. One strand involved the development and testing of the data filters on real vendor usage data. The second strand involved a survey of vendors to investigate how unique article identifiers (UAs) are currently applied. The project team considered that there was no real alternative to testing the filters developed on real data and did not feel that a survey or interviews alone would have been adequate for this purpose. Only by testing the filter with a range of vendors could we assess its viability and how generally applicable it would be. As far as determining how UAs are being applied, it was felt that a survey should include telephone interviews, as the responses were unlikely to be straightforward and there were issues that the team wished to explore further with each vendor.

The main results of the project were:

- the successful development of a UAF that will be recommended to the COUNTER Executive Committee for inclusion in the next Release of the Code of Practice. The UAF will compensate for the inflation of usage statistics, by providing a new metric – the number of successful unique article requests in a session.
- evidence that an 'unwanted html' filter would not be viable due, among other reasons, to limitations imposed by the time it takes to download an article from a browser
- an overview of current vendor implementations of unique article identifiers and recommendations to the COUNTER Executive Committee to increase consistency among vendors for the implementation of unique article identifiers.

We concluded from this project that while it is possible to further enrich the COUNTER usage data by the application of appropriate data filters, there will be limits to this in practice. These are determined on the one hand by the limitations in the data generating process, and on the other hand by the unwillingness of vendors to invest further in these processes unless there is a clear benefit in terms of the quality and value of the data thus generated.

3. Background

Project COUNTER was initiated to improve the reliability of usage statistics available for online publications. It has done so by developing Codes of Practice that set standards for the recording, reporting and delivery of vendor usage statistics. Release 1 of the COUNTER Code of Practice for journals and databases, published in January 2003, was widely adopted by vendors and the resulting usage reports were widely used by librarians.

COUNTER improves and upgrades its Codes of Practice on the basis of feedback from the library community. Release 2 of the Code of Practice for journals and databases was published in April 2005 and implemented in January 2006. This Release contains a number of improvements, including a significant enhancement of Journal Report 1: "Number of Successful Full-text Requests". In addition to providing monthly totals for successful full-text requests in all formats, Release 2 requires that a breakdown into PDF and HTML formats be provided. An analysis by Davis and Price (1) of these new data has led the authors to conclude that the design of a publisher's electronic interface can have a measurable effect on electronic journal usage statistics. Their study indicates that the ratio of PDF to HTML views is not consistent across publisher interfaces, even after controlling for differences in publisher content. They further conclude that the number of full-text downloads may be artificially inflated when publishers require users to view HTML versions before accessing PDF versions or when linking mechanisms, such as CrossRef, direct users to the full text. They propose that one solution may be to modify publisher numbers with 'adjustment factors' deemed to be representative of the benefit or disadvantage due to the interface. The COUNTER Executive Committee, however, feels that this would be difficult to implement. Davis and Price also suggest that standardization of some interface and linking protocols may obviate these differences and allow for more accurate cross-publisher comparisons. It is not the role of COUNTER, however, to specify how their interfaces should be designed or the linking protocols for services such as CrossRef. We propose instead that further research be conducted with publishers and librarians to determine whether a simple, data-processing based solution may be found. Marthyn Borghuis of Elsevier Science Direct has already investigated the application of filters to improve raw usage data in other contexts (2).

4. Aims and Objectives

The COUNTER Executive Committee agreed that the most promising remedy for any inflation of usage statistics that is due to interface effects is to work with publishers to develop an appropriate filter to be applied to the usage data. A parallel filter is already in place in the COUNTER Code of Practice to eliminate the problem of 'double clicks' by the same user on the same document; it is simple and works well. When a user clicks twice on the same HTML document within a 10-second period (30 seconds for a PDF document) this is deemed to be one request. This filter was tested by participating vendors before being included in the Code of Practice. The development of a filter to compensate for the apparent inflation of usage statistics due to interface effects would be more challenging and has implications beyond COUNTER, especially for publisher metadata. The aim of this project was to develop such a filter, assess its implications for the application of unique article identifiers, such as DOIs, and to propose how the filter could be implemented by publishers.

In the early stages of the project it was decided to test two data filters, as no experimental evidence existed to indicate which approach would work in practice. The two filters tested were:

- the 'unwanted html' filter, designed to filter out requests for full-text html articles deemed not to have been specifically requested by the user. If successful, this filter would tend to dampen the inflationary effects on usage statistics of certain vendor interface configurations.
- the 'unique article filter' (UAF), designed to provide a count for the number of requests for unique articles (irrespective of format) in a given session. If successful, this filter would tend to compensate for the inflationary effects on usage statistics of certain vendor interface configurations.

Both filters are described in detail in Section 6c below.

5. Methodology

The project was divided into three Phases, described below. The overall approach taken had two main strands, which were carried forward in parallel. One strand involved the development and testing of the data filters on real vendor usage data. The second strand involved a survey of vendors to investigate how unique article identifiers are currently applied. The project team considered that there was no real alternative to testing the filters developed on real data and did not feel that a survey or interviews alone would have been adequate for this purpose. Only by testing the filter with a range of vendors could we assess how generally applicable it would be. To determine how unique article identifiers are being applied, it was felt that a survey should include telephone interviews, as the responses were unlikely to be straightforward and there were issues that the team wished to explore further with each vendor.

Phase 1 (October-December 2005): *Development of Version 1 filter using Elsevier Science Direct and EBSCO data. Invite and brief other publishers/vendors to participate in Phase 2.* Marthyn Borghuis (Elsevier Science Direct) and Oliver Pesch (EBSCO) developed Version 1 specifications for two data filters ('unwanted html' filter and 'unique article' filter) and tested these on Elsevier and EBSCO COUNTER compliant journal usage data. It was decided to test both filters on Elsevier and EBSCO data first to eliminate unviable options and avoid involving a larger number of vendors in expending resources unnecessarily. As it became apparent from the Elsevier and EBSCO data that the 'unwanted html' filter was not viable, only the 'unique article' filter was refined further for tests on other publisher data. This process took longer than initially envisaged and was not completed until January 2006.

Also during this Phase the Project Manager approached the following vendors to seek their participation in the survey and tests involved in Phase 2 of the project: Blackwell, HighWire Press, Ingenta, Nature Publishing Group, Atypon, OUP, OVID, Springer, Wiley, American Chemical Society, American Institute of Physics and the New England Journal of Medicine. Some vendors were slow to respond and in all cases there was a discussion on the amount of work likely to be involved, as vendors had concerns about the resources and time required for the tests. While the majority indicated that they were willing to participate in a survey, only five were willing to participate in the data filter tests. The project team felt that this group of vendors, together with EBSCO and Elsevier, would provide a representative cross section of current online journal vendors, as they ranged from an aggregator and the largest journal publisher at one end, through to smaller publishers with very heavily used journals.

Phase 1 was not completed until February 2006.

Phase 2 (January-March 2006): *Test viability of Version 1 filter with key major publishers/vendors. Assess whether modifications will have to be made to current publisher practice regarding the application of unique article identifiers.* Angela Conyers/Pete Dalton, together with Peter Shepherd, designed and conducted a questionnaire-based telephone survey to obtain feedback from publishers on current practice regarding the implementation of unique article identifiers. This aspect of the project was completed on schedule.

It took longer than initially envisaged to set up the vendor tests of the UAF, mainly owing to difficulties vendors had in scheduling the work. These tests began in mid-March and were due to be completed by the end of April. This delayed completion of this aspect of the project from mid-March to the beginning of May. It was felt by the project team, however, that full tests of the data filter on real, COUNTER-compliant usage data, were necessary in order to validate the filter as being generally applicable.

Phase 3 (March-May 2006): *Analysis of survey results, assessment of viability of filter, preparation of report for JISC, identify relevant modifications to the COUNTER Code of Practice.* Peter Shepherd, Angela Conyers and Pete Dalton analysed the results of Phase 2

and discussed this analysis with other team members prior to drafting the final report to JISC and proposing modifications to the COUNTER Code of Practice.

6. Implementation

This project had three main objectives:

- a. The development of a filter to be applied to usage data that would dampen or compensate for the inflationary effects of certain vendor interface configurations.
- b. An assessment of current vendor practice regarding implementation of unique article identifiers, such as DOIs
- c. Propose how the filter could be implemented by vendors and recommend appropriate modifications to the COUNTER Code of Practice to the COUNTER Executive Committee.

In the first project team meetings held in October and November 2005 our focus was on the development of the technical model for the data filters and their testing on Elsevier and EBSCO data. In the course of these discussions we concluded that the best approach to tackling this challenge is to test two distinct data filters, each of which would give a different insight into usage. These are:

- The 'unwanted html filter': this filter is based on the assumption that the time that elapses between a request (click) for an html full-text journal article and the next request (click), as well as the nature of the next link, is a measure of the value of that html document to the user. In other words, one can set a time filter that can be applied universally to eliminate 'unwanted' html requests – usually due to certain publisher interface configurations- from the COUNTER usage statistics. The Elsevier and EBSCO tests were designed to enable us to estimate the optimal timing of this filter. The methodology for the development of this filter is described in Appendix A.
- The 'unique article filter': this filter is based on the assumption that we can use a unique identifier for an article (irrespective of format) and an identifier for a session to derive the number of unique article requests per session. This filter does not eliminate 'unwanted html' usage, but would provide another level of insight into usage. The technical model for this filter is included in Appendix B.

In both cases usage data compliant with Release 2 of the COUNTER Code of Practice for journals and databases were used in the tests.

DEVELOPMENT AND TESTING OF THE 'UNWANTED HTML FILTER'

This process began in November 2005 and continued into January 2006, using the methodology described in Appendix A. Following the initial tests it became apparent that it was going to be difficult to set a time filter to eliminate 'unwanted html' requests as described above. Both EBSCO and Elsevier tested the time filter over a wide range of intervals. EBSCO (Appendix C) found that if a data filter is implemented after a time interval from the initial html request of 2 seconds, only 4% of full-text html views will be eliminated; this increases to 6% after 6 seconds and to only 7% after 8 seconds. EBSCO tested the filter in two different scenarios: first in 'regular sessions' where the user begins the session in an EBSCO environment; second, in 'link-in' sessions, where the user comes to the journal article via a link from another site or service. The results were very similar for both scenarios. Elsevier observed a similar pattern in their data. The picture is further complicated by apparent large variations in interface/browser performance. There can be differences of several seconds in the time it can take different browsers or interfaces to render an html document. There can be

a variety of reasons for this, but it means that data filters in the <10 second range appear to be unsatisfactory in terms of filtering out 'unintended' usage of full-text html. Setting the time filter at a longer interval presents other problems. The results from both EBSCO and Elsevier show that if a 30 second filter is implemented, around 60% of full-text html counts are eliminated. In the case of EBSCO this occurs whether the 'auto-html' facility, which displays the full-text html article whether or not the user has requested it, is on or off. We feel that it is not realistic to assume that the majority of html requests that are still open after 30 seconds are unwanted. Our results indicate, therefore, that specifying a time filter designed to eliminate 'unwanted html requests' will not be a practical way of mitigating the inflationary effects of certain interfaces on COUNTER usage statistics. If a short time filter (<10 seconds) is implemented the time taken to render the documents becomes significant, undermining the function of the filter. If a longer time filter is set, too high a proportion of the html requests are eliminated. Given the consistency of the EBSCO and Elsevier results in this respect, we decided that it would be pointless to ask a larger number of vendors to allocate resources to testing this filter on their own data.

Initial results on the application of the 'unique article filter' appear to be more promising. This would make viable a new metric 'number of unique article requests' in a single session a practical metric for the publisher to record. We already know from previous studies that this is a metric that many librarians would like to have.

DEVELOPMENT AND TESTING OF THE 'UNIQUE ARTICLE FILTER'

EBSCO and Elsevier data

Effort focussed on the 'unique article filter' began in December 2005, once it became apparent that the 'unwanted html filter' was not going to be viable. Oliver Pesch and Marthyn Borghuis developed the specification for this filter (Appendix B) and tested it on Elsevier and EBSCO data during January and February 2006. The Elsevier analysis was based on 12.2 million sessions, covering 15.7 million full text articles in PDF or HTML. The EBSCO tests were conducted using 27,000 sessions representing over 800,000 separate activities in the transaction logs.

As in the case of the 'unwanted html' filter, EBSCO tested the filter in two different scenarios: first in 'regular sessions' where the user begins the session in an EBSCO environment; second, in 'link-in' sessions, where the user comes to the journal article via a link from another site or service. The results are summarised in Appendix D. Long-duration sessions were excluded from the results to ensure that a session is truly a user session.

On the graphs in Appendix D the left-hand number represents the actual filter effect, i.e. the total percentage by which the total count is reduced by the unique article filter. In the case of Regular Sessions this reduction is 28%. In other words the unique article count is 28% less than the html+PDF count. (Note: In order to make the volume of data to be processed more manageable EBSCO also tested the effect of chopping the session up into time increments.)

These results indicate that the UAF has considerable merit as it does eliminate double counting from multiple format requests; the 28% reduction observed by EBSCO would appear to be reasonable. Applying the same UAF to Elsevier usage data resulted in a lower reduction (21.9%) in the total html+PDF count than observed by EBSCO. See Table 1 below:

Table 1: Results of UAF tests on Elsevier usage data

Sessions with Unique Articles (PDF or html)	% of total	Cumulative %
PDF	49.1%	49.1%
html	11.4%	60.54%
One full text article all other sessions	17.6%	78.1%

The effect of the UAF on Elsevier-Science Direct data is to reduce the count to 78.1% of the total; i.e. a reduction of 21.9%. This includes all sessions, without time slicing.

It is clear that the application of the UAF has a significant effect on both the EBSCO and Elsevier usage data. In both cases the reduction in counts is substantial. On the other hand, there is a 6% difference in the impact of the UAF on the EBSCO and Elsevier data. This is not, of itself, a problem; indeed, there is no reason to expect that the reduction would be the same in both cases. Not only is the Elsevier and EBSCO content different, but it is presented in a different context and in a different interface.

Other publisher data

It was a major challenge not only to persuade other publishers to devote time and resources to running the unique article filter tests, but also to share the resulting data with the project team. Of the 11 publishers who agreed to participate in the 'Unique Article Identifier' survey, only 5 agreed to carry out the tests according to the specification in Appendix B. They are identified as Publisher A, Publisher B, Publisher C, Publisher D and Publisher E. The data recorded are provided in graphical form below, while the underlying data are provided in the spreadsheets contained in Appendices E to I.

Summary results for Publishers A to D, showing the percentage reduction in the total count resulting from the application of the Unique Article Filter

Figure A

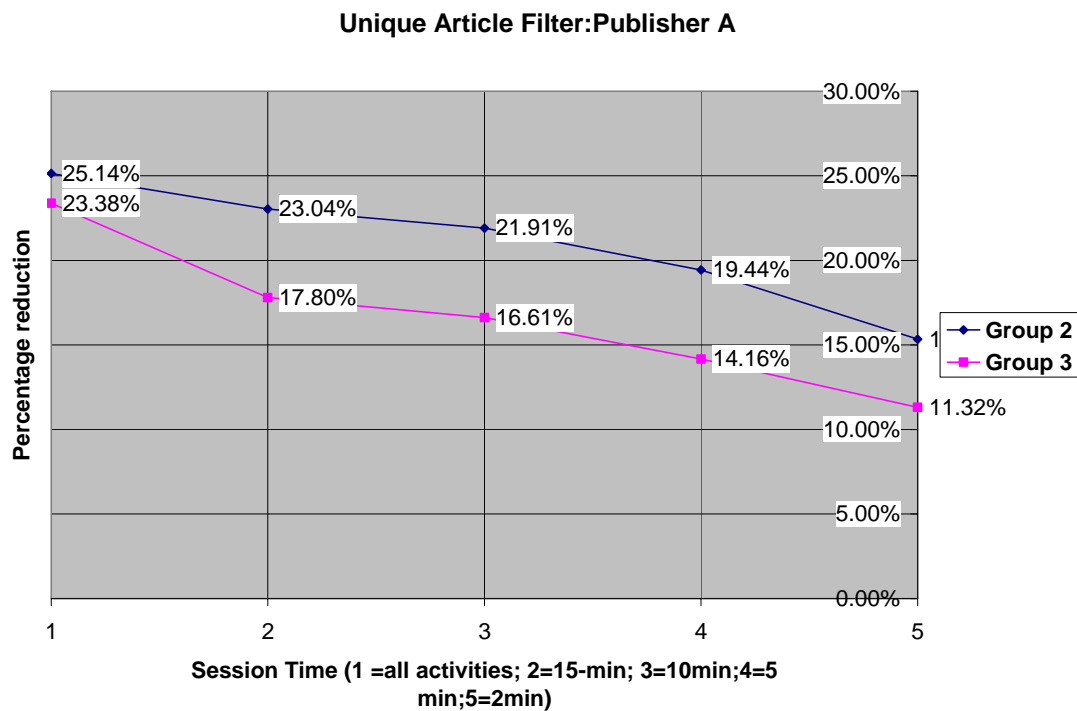


Figure B

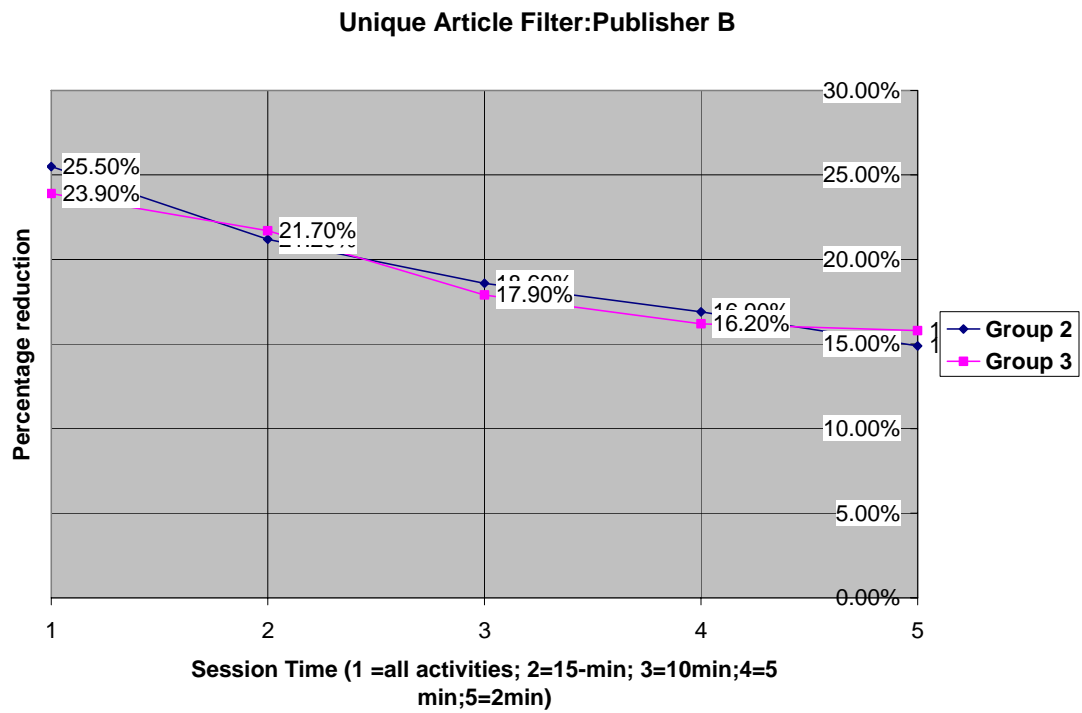


Figure C

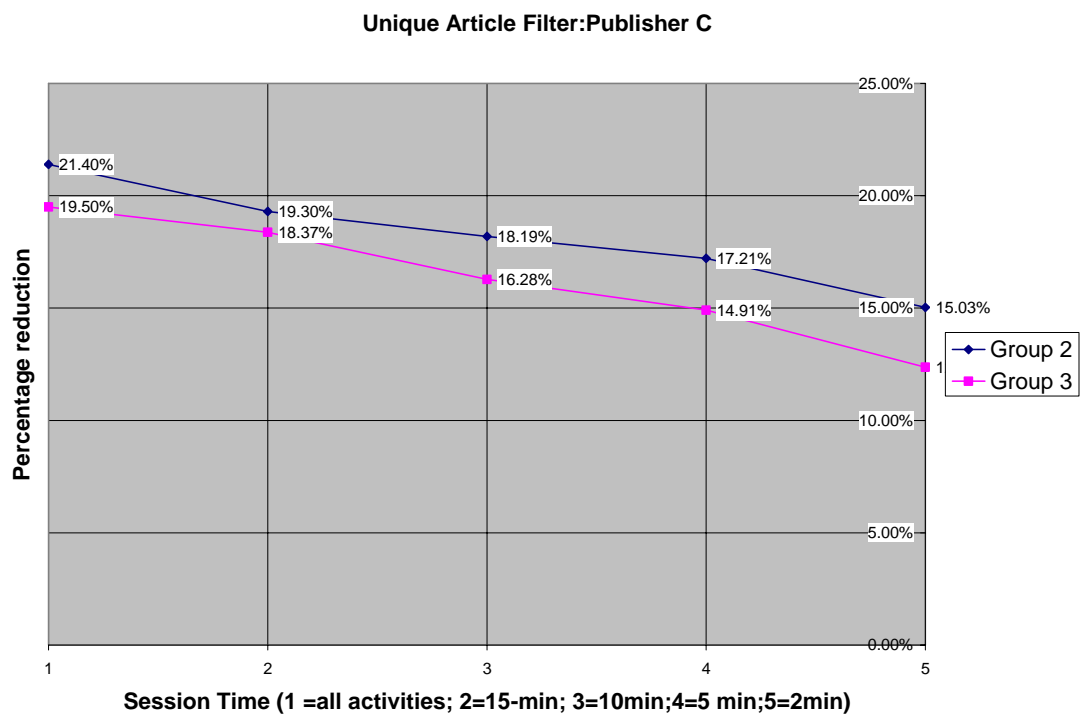


Figure D

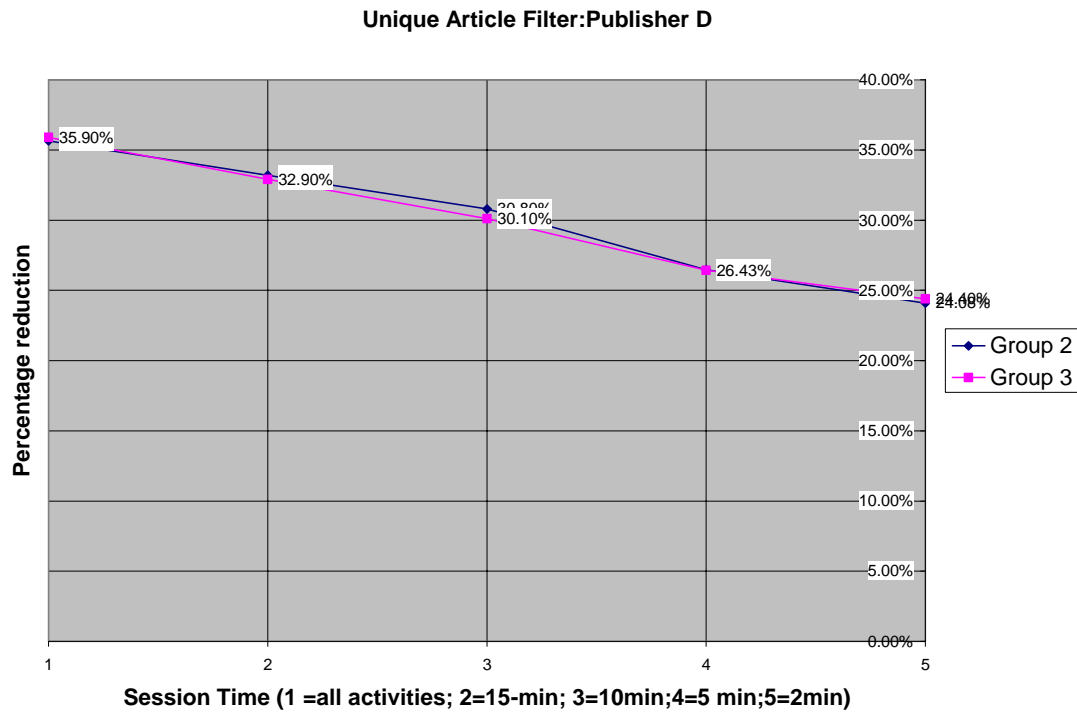
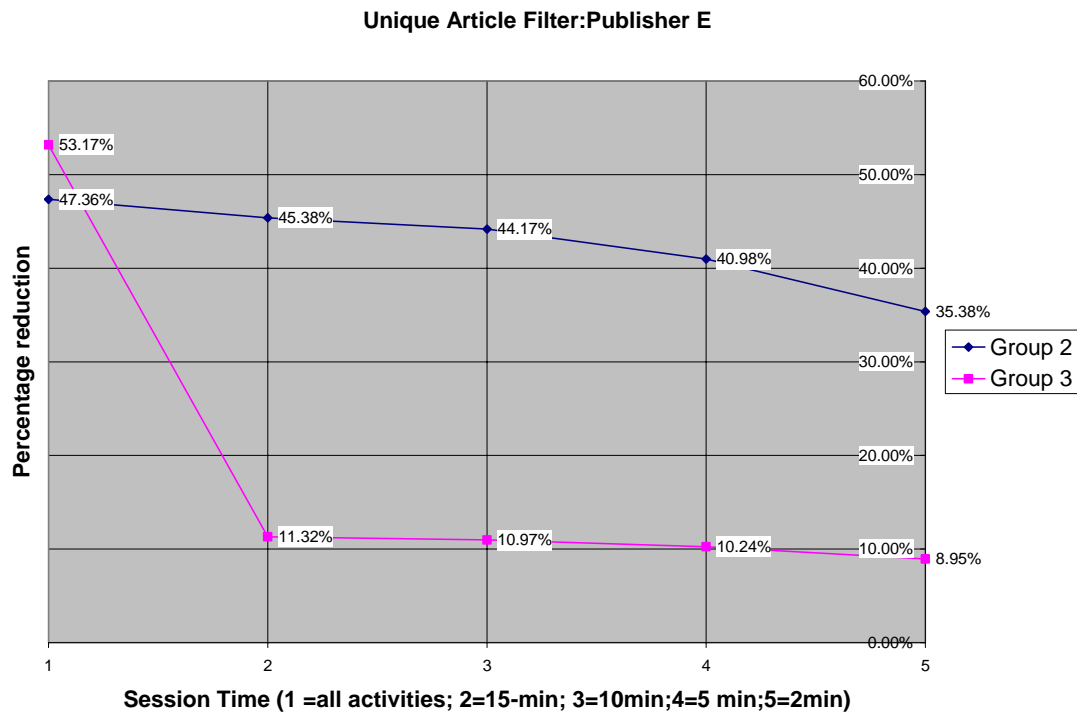


Figure E



Note:

1. Group 2 contains sessions with 2 to 10 full text requests within a session
2. Group 3 contains sessions with greater than 10 requests (this group will most likely represent multiple users sharing a session, or crawlers pulling full text).
3. In order to make the volume of data to be processed more manageable publishers also tested the effect of chopping up the session into time increments. (If a vendor is uncertain

about their sessions and is unsure whether they represent multiple users -eg long sessions- then slicing the session into defined windows could mitigate this problem). The data indicate that setting a session time limit of 10 minutes changes the UAF reduction by 1%-7% in the majority of cases. Publisher E, Group 3 data, however, shows a very different pattern and indicates that further tests will be required before this approach can be validated.

Table 2 below summarises the results of the full-session tests by Publishers A to E.

Table 2: Effect of Unique Article Filter on Full-text Article Count

Vendor	% Reduction (Group 2)	% Reduction (Group 3)	ratio PDF/html (Group 2)	ratio PDF/html (Group 3)
Publisher A	25.14%	23.38%	0.64	0.57
Publisher B	25.50%	23.9%	4.00	4.50
Publisher C	21.40%	19.50%	7.69	6.80
Publisher D	35.65%	35.90%	1.05	0.96
Publisher E	47.36%	53.17%	0.97	0.77

Note:

1. Full session data only; for full data sets, including sessions to which a time slice has been applied, see Appendices E-I.

The results for Publishers A to E confirm the significant reduction in the total usage count already observed for the EBSCO (28%) and Elsevier (21.9%) full-session data when the UAF is applied. The range in the percentage reduction for the five publishers is considerable, at 20%-53%. While such a spread is not unreasonable, it is not clear from the data what factors contribute to it. Apart from interface configurations, differences in patterns of full text article usage between disciplines may be significant, but the data collected for this study is insufficiently granular to yield insights into such phenomena. Either could significantly affect the relative proportions of PDF/html downloads. Where the PDF/html ratio is the highest one might expect the UAF to produce the smallest percentage reduction in total count, assuming that multiple downloads of the PDF version of the same article are much less likely than multiple downloads of the html version. Table 2 shows that while the lowest percentage reduction observed is for Publisher C, where the PDF/html ratio is highest, the lowest PDF/html ratio (Publisher A) does not correspond with the highest percentage reduction, which is seen with Publisher E. There is no clear pattern.

The conclusions one can draw from the EBSCO and Elsevier tests, as well as the data presented in Figures A-E and in Table 2 above are:

- the UAF can be successfully implemented by a range of publishers, removes non-unique article requests in a session and results in a significant drop in the total count for the 'number of successful full-text requests' for all publishers
- the percentage reduction ranges from 20%-53% for a full session of unlimited duration. Given the differences in publisher interfaces, the subject content covered, the formats used and the context in which that content is available, this range is not surprising, but there is no clear pattern in the data presented here that explains it fully.
- the results for the Group 2 and Group 3 tests are hugely different only for Publisher E. For the other four publishers the UAF is not significantly affected by whether the number of successful full-text requests within a session is greater or less than 10. It is not clear from the data why the effect should be so different for Publisher E.
- the application of time slices at different periods (Figures A-E) reduces the effect of the UAF in a broadly similar pattern for Group 2 and Group 3 data for publishers A-D and for the Group 2 data for Publisher E. The Group 3 data for Publisher E, however, shows a markedly different effect once the time slices are applied to the data; the reduction in the total count falls from 53% to 11%. There is no clear reason for this in the data, but it indicates that applying time slices to the data may distort the effect of the filter in some cases.

SURVEY OF VENDORS REGARDING CURRENT PRACTICE REGARDING IMPLEMENTATION OF UNIQUE ARTICLE IDENTIFIERS

The following 11 publishers and vendors took part in a questionnaire survey which was conducted by Angela Conyers (Evidence Base) in the period 16-24 March 2006.

ACS
AIP
Atypon
Blackwell
EBSCO (email)
Elsevier (email)
HighWire
Ingenta
MPS
OUP
Wiley (email)

A list of questions and a brief background paper (Appendix J) were emailed to contacts in advance and this was followed up with a request to book a time for the researcher to phone. Eight out of the 11 were interviewed by phone; the remaining 3 provided email responses. The full results are provided in Appendix K. The responses to each questionnaire summarised below:

1) Do you attribute the same Unique Article Identifier to a particular full-text journal article, irrespective of format (PDF, html, etc)?

All but one vendor answered YES to this question. AIP used separate unique article identifiers (UAs) for PDF/Postscript files and for HTML, with PDF/Postscript files retaining their identifier and HTML being assigned its UAI during the COUNTER process.

2) Is this then a permanent attribute of that article?

Although all vendors apart from one (HighWire) gave a definite YES to this question, answers to the following question indicate that in the case of Ingenta also the UAI was not a permanent identifier as it changed from the pre-print to the publication stage.

AIP reported that their identifiers were permanent once attributed, but as indicated in their reply to question 1, different identifiers were used for PDF/PS and HTML formats and those for HTML were not applied until the publication stage.

3) If so, do you use DOI for this purpose? If not, what is your current practice?

5 vendors (ACS, Atypon, Blackwell, Elsevier and OUP) were using DOI.

One (Wiley) was using DOI and also an internal identifier, the 'Object Identifier' OID (as the OID was already in use by the Wiley Interscience electronic publishing system before the DOI was invented). There is a one-to-one relationship between the OID and the DOI. The OID is assigned when an article is cleared for publication, continues to be used as an internal identifier within Wiley, is used to generate the article URL, and is used for all counting purposes. The equivalent DOI is shown on the article itself and is used in CrossRef.

One (MPS) reported that some of their publishers use full DOI, but most do not; some use part of the DOI after the publisher prefix, since full stops and slashes in URLs cause problems.

Four (AIP, EBSCO, HighWire* and Ingenta) were using an in-house or proprietary system generally based on vol, issue, page

*HighWire reported that they used vol,issue,page as primary identifier, as usage of DOI varied between their publishers.

4) At what stage in the publishing process do you apply the unique article identifier? Are different article identifiers applied at different stages of the publishing process e.g. pre publication and publication.

7 vendors (ACS, Atypon, Blackwell, EBSCO, Elsevier, MPS, OUP) reported that the UAI remained the same from the stage the article was made available externally, i.e. at the pre-print stage. Some mentioned using a different number for internal tracking before that stage .

One (Wiley) assigned the DOI when the article is made publicly available and the DOI is the only article identifier which is seen externally online and in print. The equivalent OID, for each article continues to be used internally. Wiley do not offer pre-prints.

Two vendors (HighWire and Ingenta) used a different identifier for pre-print articles. In the case of HighWire this was a unique identifier which was mainly but not necessarily the DOI, depending on the publisher. When the article is published, it gets a volume/issue/page/identifier which is used for the COUNTER report. Reports to publishers use the volume/issue/page/identifier, except for ahead of print usage which is based on the DOI or similar. This was explained as follows:

A few of the publishers we work for chose not to join CrossRef but they have pre-print articles. Essentially what they do is use an identifier that is DOI-like (the same format that is used for DOIs) but it is not registered with CrossRef. So, it's not officially a DOI but the HighWire system treats it as if it is and all of the associations between it and the volume/issue/page identifiers work as we described.

Ingenta used the DOI for preprint, then changed to a proprietary number on publication. In response to a follow-up question on the counting of preprint article usage, this explanation was provided by Ingenta:

At the moment, we're not able to count usage of pre-print articles due to the legacy system on which they still reside. This is due to change when we consolidate all our content in a new metastore over the course of the next few months; at that point, we'll also have a single identifier for any article throughout its life with us. We will then need to rewrite our stats so it will be a little while before we're offering statistics for pre-print articles -- but when we do, we'll be able to transition stats from pre-print through to "normal" article status, such that usage for a given article can be comprehensively and accurately counted.

One vendor (AIP) used the same identifier for PDF/Postscript articles from the preprint stage, but did not assign the HTML identifier until after publication during the COUNTER process. The article identifier is a 30 character combination called the CVIPS. PDF/PS articles contain this value in the URL. The HTML article contains the same format and the article identifier can be derived from the HTML URL.

5) In which formats do you publish full-text journal articles?

All published in PDF and HTML. Other formats mentioned were:

- Postscript (AIP)
- PDF with references (Atypon)
- DTD controlled XML (Blackwell)
- Many variations of PDF including translation (HighWire)

Audio, video, Excel, TIF (MPS)
PDF – low and high resolution (OUP)

HighWire commented on the large number of variations of the PDF format in which they published.

6) Are you able to identify activities for a given user session within the usage logs?

8 vendors replied YES. Two others indicated that it was technically possible, ACS saying they 'don't drill down that far' although the data is available and HighWire that it was not used in their usage reporting. One (OUP) explained that they could not currently do this with the web server log files provided by HighWire from which they generated their own statistics. [Note: HighWire explained that this information was in the logs provided to OUP but not currently used].

7) How do you identify the user session?

Apart from OUP, vendors were all able to describe means of identifying user sessions. These ranged from:

- IP address alone (ACS)
- IP address with user agent (combining operating system version number of browser version number) – MPS (IBM SurfAid system)
- Unique session ID (Atypon, Blackwell, EBSCO, Ingenta), used with a combination of account, cookie ID, IP address, user cookie, registered user (AIP, Elsevier)
- Log including customer ID (HighWire)
- Session cookies (Wiley)

8) Would it be a problem for you if COUNTER specified a new report requiring that the number of unique article requests in a given session be counted, in which the DOI (or another stable identifier) was required as a unique article identifier, to provide a permanent attribute for each full-text article, irrespective of format?

All vendors acknowledged that it would be technically possible to produce this new report but raised the following issues:

- How to define sessions (ACS, Elsevier, MPS, OUP)
- Concern over costs and extra work involved, including programming costs and ongoing processing costs (AIP, HighWire, Wiley)
- Failure to measure added value of different views (HighWire)
- Confusion to customers from having another set of reports (Wiley)
- Effect of e-journal platform design (Elsevier, Wiley)
- Costs v benefits (AIP, MPS)
- Need to have this extra report (AIP, HighWire)
- May delay production of Counter reports (HighWire)

Six vendors (Atypon, ACS, Blackwell, Ingenta, MPS and OUP), although recognising that some changes may be involved, felt that this new report would not be a major problem for them if required for COUNTER compliance, providing session definitions were clear. This would apply also to EBSCO and Elsevier, as partners in the project.

Three others (AIP, HighWire and Wiley) had more serious reservations regarding likely extra costs and possible benefits.

All were willing to look at the test requirements, though stressed that their ability to take part would depend on timing and the amount of work involved.

It is clear from the results of this survey that there is a wide range of practice among vendors in terms of the application and implementation of UAIs. Particularly noteworthy are the following:

- all but one of the vendors attribute the same UAI to a particular full-text journal article, irrespective of format. The exception is AIP.
- most vendors use either the full DOI, or a subset of the DOI, as the UAI, but a significant minority (4) use an in-house or proprietary system
- the majority of vendors reported that the UAI remained the same from the stage at which an article becomes available externally, i.e. at the pre-print stage. HighWire Press and Ingenta use a different identifier for pre-print articles. Interestingly, Wiley use the DOI for the 'traditional' print article, but another identifier for the same article online.
- the majority of vendors are able to identify activities for a given user session within the usage logs
- the majority of vendors would be able to provide a new COUNTER report on the number of unique article requests in a given session. Some have reservations about the cost/benefit ratio involved in generating this report.

The wide range of practice among vendors on the application and implementation of UAIs appears to be a result of both history and current requirements. In some cases, such as Wiley's using different identifiers for the print and online versions of the same article, current practice does not appear to be logical. There would be clear advantages for the industry as a whole to have more standardisation in the application and implementation of UAIs. From the usage statistics perspective it would make sense for there to be some 'best practice' guidelines, covering the following issues:

- Article formats and UAIs: a unique full-text article should have the same UAI applied to all formats, whether in print or in electronic form
- Stage at which UAI is applied: the UAI should be applied no later than the point at which the full-text article is first publicly available. It should then be a permanent attribute of the article.
- Preferred UAI: the preferred UAI is the DOI, but others are acceptable, provided they meet the two criteria above.

7. Output and results

The two main outputs of the project are:

- a. A data filter that can be applied to the data contained in the COUNTER Journal Report 1 (Number of successful full text article requests), which will result in a meaningful figure for the number of successful unique article requests in a given session. This data filter looks promising for general applicability to COUNTER-compliant journal usage data.
- b. An overview of current vendor practice in the implementation of UAIs, and a set of recommendations for best practice in the implementation of such UAIs.

These outputs are consistent with the original objectives of the project. While a filter that eliminated 'unwanted html' full-text requests proved unviable as a means of mitigating the inflationary effects of certain vendor interface configurations, the unique article filter will provide users with another metric – number of unique article requests in a given session – that will provide them with an alternative perspective on usage that will help them assess the value of a given online journal. If implemented in COUNTER, users will have the following data on journal usage:

- total number of successful full-text article requests (in all formats)
- total number of successful full-text article requests in html format
- total number of successful full-text article requests in PDF format
- total number of unique article requests in a given session

The result is a further enrichment of the COUNTER data, which will allow librarians to compare usage in different formats across different sets of publications, between publishers, within disciplines, etc., enabling them to assess the value of these formats in a variety of contexts and to identify patterns of usage that merit further investigation. It will also allow them to build on the methodology developed in the JISC-sponsored survey of online usage of journals included in the NESLi 2 deal (3). Metrics such as 'Cost per Use' could be extended to cover 'Cost per Unique Article Use'.

8. Outcomes

This project is the first to test data filters on usage statistics from different publishers. This would not have been possible pre-COUNTER and we have explored new territory. We have also learned some important lessons about the practical challenges involved in attempting to extract more information from vendor usage data. In addition to the outputs listed in Section 7 above, there are two broad conclusions to be drawn for this project. The first is based on the results of the project, the second on the methodology followed in the project:

- a. Conclusions based on the results: it is clear that, while it is possible to further enrich the COUNTER data by the application of appropriate data filters, there are limitations to this in practice. These are determined on the one hand by the limitations in the data generating process, and on the other hand by the unwillingness of vendors to invest further in these processes unless there is a clear benefit in terms of the quality and value of the data thus generated. Also, there comes a point where the additional cost of generating more detailed online usage data is not justified by the benefits to vendors or librarians and other methods of obtaining insights into online usage (surveys, interviews, etc.) may be more effective. It is important to see usage data in this wider context and to avoid generating ever more granular data because it is possible to do so.
- b. Conclusions based on the project methodology: the rate-limiting step in the progress of this project has been the time and resources that vendors are willing to schedule for this kind of research. There is a core of vendors that have a genuine interest in and commitment to usage-based research, while others have more problems in devoting resources to this type of work, which is not necessarily regarded as being core to the business. In retrospect the original timetable for the project was too ambitious and it would have been more realistic to extend it by 2 months (to mid July 2006) to take into account vendor scheduling problems.

9. Implications

This work has the following implications for COUNTER and for the wider community

- a. Further improvements to the COUNTER Code of Practice: a core COUNTER philosophy is to strike an appropriate balance between the demand of librarians for more reliable online vendor usage data and the ability or willingness of vendors to bear the additional costs of recording and reporting such usage data. The project team believes that the UAF developed in this project strikes the right balance and will recommend that the COUNTER Executive Committee includes this filter in the next Release of the Code of Practice.
- b. More systematic use of COUNTER-compliant usage statistics in a wider market research context: the new tools proposed in this report enhance the capability of vendors to generate more granular COUNTER usage statistics. As the body of COUNTER-compliant usage data grows it becomes increasingly viable to use this to monitor trends in usage at institutional and supra-institutional levels. To do so effectively will require further investment in

infrastructure and skills. JISC is already investigating the feasibility of developing a central UK repository for online usage data. If this does proceed it will not only form a basis for more systematic use of COUNTER data in the UK, but will also be a model for such initiatives worldwide. If it is concluded that such a central repository is not viable alternative approaches to promoting the systematic analysis of usage data will have to be developed. These could include JISC developing a set of tools to be used by librarians and a greater focus on improving individual librarian skills in this respect.

- c. Greater standardisation among vendors on the implementation of UAIs: while this report recommends a Best Practice for the Implementation of UAIs by vendors it does so only in the context of enhancing online usage statistics by effective implementation of the UAF proposed here. The implementation by vendors of UAIs has implications far beyond the requirement for more reliable usage statistics, however, and a more consistent approach would have a number of wider benefits. For example, in an online environment in which e-prints of articles are available in advance of formal publication and in which the same article is available in print, as well as in a variety of electronic forms, more consistency among vendors as to which UAI to use and at what stage in the publishing process it should be applied is desirable. A broader discussion should be initiated by JISC to agree best practice for the application of UAIs, with a view to developing an international standard. The recommendations listed at the end of Section 6 above are important from the COUNTER perspective and should be taken into account in a wider discussion.

10. Recommendations

The recommendations of the project team are as follows (they are relevant to the research community rather than to the teaching or learning communities):

- a. To COUNTER: implement the proposed UAF in the next Release of the Code of Practice. This filter should be applied to full sessions, without time slices, until the effects of time slicing the sessions has been more fully investigated.
- b. To COUNTER: in view of the failure to develop a successful 'unwanted html' filter, COUNTER should provide guidelines on 'best practice' for vendor interface design from the perspective of delivering reliable usage statistics. These guidelines should expressly state that vendor interfaces which require the user to download the html version of a full-text article before they can download the PDF version are not best practice.
- c. To COUNTER: include the following Guidelines for implementation of UAIs be incorporated into the next Release of the Code of Practice:
 - Article formats and UAIs: a unique full-text article should have the same UAI applied to all formats, whether in print or in electronic form
 - Stage at which UAI is applied: the UAI should be applied no later than the point at which the full-text article is first publicly available. It should then be a permanent attribute of the article.
 - Preferred UAI: the preferred UAI is the DOI, but others are acceptable, provided they meet the two criteria above.
- d. To JISC: initiate a wider discussion on best practice for the implementation of UAIs by vendors. Include in the discussion representatives of the vendor, library and standards communities.

11. References

1. Davis, PM and Price, JR, "eJournal interface can influence usage statistics: implications for libraries, publishers, and Project COUNTER" (accepted for publication in JASIST). Available at: <http://people.cornell.edu/pages/pmd8/interface.doc>
2. Borghuis, Marthyn, "What to COUNT and What Not? A White Paper on the filters to be applied to raw usage data before usage analysis can start." Elsevier, 2004.

Available at:

http://www.info.sciencedirect.com/librarian_help/usage_reports/sd_white_paper_2004_02.PDF

3. Conyers A., and Dalton P., NESLi 2 Analysis of Usage Statistics
(<http://www.ebase.uce.ac.uk/docs/jiscnesli2summaryeb.PDF>)

12. Appendices

Appendix A: html filter enhancement methodology
Appendix B: Unique article filter testing methodology
Appendix C: EBSCO results for 'unwanted html' filter tests
Appendix D: EBSCO results for unique article filter tests
Appendix E: Publisher A results for unique article filter tests
Appendix F: Publisher B results for unique article filter tests
Appendix G: Publisher C results for unique article filter tests
Appendix H: Publisher D results for unique article filter tests
Appendix I: Publisher E results for unique article filter tests
Appendix J: Unique article identifier briefing paper
Appendix K: Unique article identifier survey results