Keystroke Economy: A Study of the Time and Effort Involved in Self-Archiving

Leslie Carr, Stevan Harnad
University of Southampton

Abstract

A common objection to self-archiving is that it is an extra task that puts an unnecessary burden on each researcher. In particular, the need to enter the extra bibliographic metadata demanded by repositories for accurate searching and identification is presumed to be a particularly onerous task. This paper describes a preliminary study on two months of submissions for a mature repository and concludes that the amount of time spent entering metadata would be as little as 40 minutes per year for a highly active researcher.

Background

Open Access Archiving (OAA) is a mechanism for making scientific output (papers or articles) accessible as a parallel supplement to the usual scientific publication process. It is accomplished by depositing a copy of the published work in an Open Access Archive (for example, in an institutional or subject-based repository). In this paper, the terms "archive" and "repository" are taken as equivalent. *Archive* has historical precedence in this context – for example, *the Open Archives Initiative* and the 'self-archiving initiative'; *repository* is a more recent term intended to avoid irrelevant connotations of bulk storage and preservation for documents whose primary utility is "archival" rather than the immediate access and usage that is the primary rationale for Open Access.

In the self-archiving model, authors provide OA to their own research output by depositing it into an OA archive. As part of this process, the author (or designee) must upload a copy of the paper and also enter some simple metadata (author, title, publication, date, *etc*.) that describes the paper, making it interoperable with papers self-archived in other OA archives, and allowing the metadata to be harvested, citation-linked, and searched seamlessly as if all papers were in one global archive.

The metadata are hence important for maximising the semantic interoperability and the power of navigation, analysis and retrieval over OA archives (citation searching, bibliographic matching). Entering the metadata is, however, an extra task, over and above merely depositing the paper's text, and one that is frequently omitted by authors because it has the reputation of requiring hard work (or at least extra time and effort that must be clearly justified).

This reputation of being a time-consuming chore has been a significant disincentive for individual self-archiving [3] and hence an impediment to achieving Open Access on a global scale.

In order to investigate the actual time and effort involved in self-archiving we instrumented the user interface of a mature research repository and collected timing data for researchers' deposits over a period of several months.

The Repository

The chosen repository is the EPrints-based research archive of the School of Electronics and Computer Science at the University of Southampton. The repository has been in use since 2001 (with a non-full-text publication database in place prior to that date). Since 2003, it has been mandatory to deposit all research article output in the repository, which has been used as the database for all administrative research returns since that date. The repository contains almost 9,000 publications, deposited by about 150 researchers in research areas ranging from Electrical Power applications, through Microelectronics and Software Engineering to Digital Libraries. All records contain bibliographic metadata, but currently only 23% also include the full text. This is mainly because of the legacy backlog of several decades of papers and partly because of the delay in uploading 'final version' postprints. The amount of full text deposited rises to 50% in 2004, but this still reveals a less-than-complete compliance with a School Policy that mandates the use of its repository. The reasons for this include include individual authors' legal wories about copyright and publication contracts and the pragmatic concern about the extra effort involved in depositing material into an archive [1].

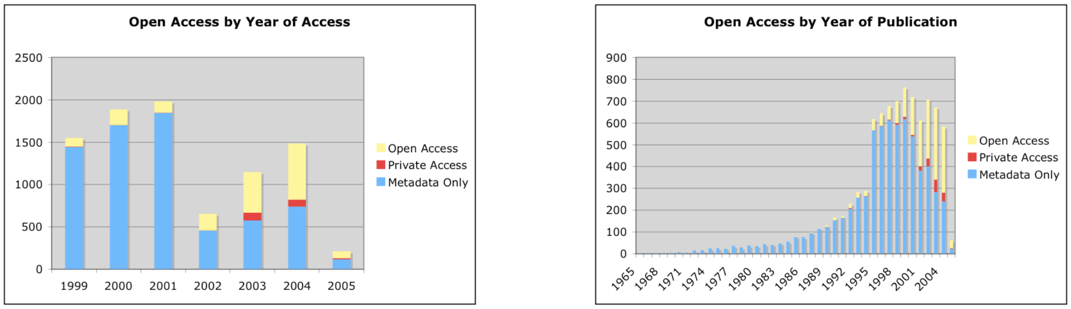


Figure 1: Open Access Deposits by Year of Deposit and Year of Publication

The repository is built on GNU EPrints software version 2.3, configured for tight integration with the School's management and administrative processes (e.g. using the school's user authentication and providing lists of publications for users' home pages and automatic lists of recent publications for marketing purposes).

The contents of the School's repository will eventually be migrated to the University's EPrints server [3], but currently it is an independent application that is configured for the local requirements of the ECS research community.

In particular, the metadata gathering interface is concentrated into a single Web form (which was the default configuration in the EPrints software prior to version 2.3) and not split between a number of forms on individual web pages in a planned workflow, as is common in more recent implementations [4].

The paper deposit workflow is shown in figure 1: first a paper type is chosen, then all the paper metadata are filled out, then the files which constitute the full text are uploaded (*e.g.* Microsoft Word, PDF or HTML). Lastly the record is displayed for verification by the author before finally being deposited in the archive.

For the purpose of this investigation, logging code was added to the scripts of the deposit forms so that every new web page left an entry in a log file recording the user, the eprint id, the web page requested and the button which was pressed in order to reach that page, along with a timestamp. Logging data were collected for 67 days from December 2004 to February 2005, resulting in deposit and editing sessions for 260 new eprints. (Other sessions that represented modifications to already-existing eprints were ignored.)

```
Timestamp     EprintID    User    Page            Button Press
1106833807    10390       11      type            new
1106833811    10390       11      meta.default    next
1106834121    10390       11      files           next
1106834123    10390       11      docmeta         newdoc
1106834132    10390       11      fileview        next
1106834154    10390       11      fileview        upload
1106834157    10390       11      files           finished
1106834160    10390       11      verify          next
1106834170    10390       11      done            submit
```

Figure 2: Log file for EPrint ID 10390

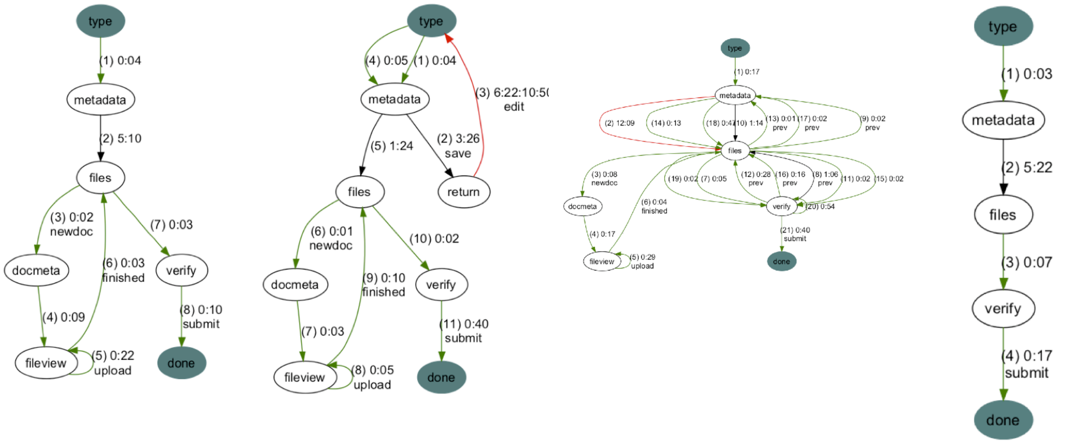
Sample Results

A visual display was developed to help analyse by sight the deposit behaviour for the captured sessions. Figure 3(a), which displays the log file in figure 2, corresponds to the normal form of a deposit. Each node represents a separate web form (i.e. web page) in the deposit process, and each arc represents the transition between two pages. The arcs are labelled with a number indicating the order in which the pages were visited plus the number of minutes and seconds that elapsed before the transition took place (corresponding to the amount of time that was spent on the previous page). Shaded nodes represent the beginning and end of the deposit process. Transitions that took longer than ten minutes are colored red. Very long delays are indicated in the form days:hours:minutes:seconds.

The session represented by figure 3a begins with the user specifying the *type* of eprint (journal article, workshop presentation *etc*.) This stage determines the set of metadata items that are relevant for the next page (*e.g.* journal volume and issue numbers *vs* workshop location), and is labelled *metadata* in the displays. The arc between the two nodes in figure 3a shows that four seconds elapsed between the user's request for the *type* and *metadata* pages, which is interpreted as the total time taken to determine the correct choice, press the appropriate button and click on the "Next Page" button. Also included in this figure are delays in transmission through the network, delays in interpretation and display on the browser and user task interruption by phones, email, visitors, coffee *etc*. The next arc in figure 3a shows that five minutes and ten seconds elapsed while the user filled out the metadata fields and before they requested the next page that begins the process of uploading the various documents associated with this eprint. Here there are several substeps (a) providing document format metadata on the *docmeta* page and (b) uploading at least one file for each format on the *fileview* page – returning to that same page for the opportunity to upload more files forms the loop labelled '(5)'. Here the delay of twenty-two seconds is caused by the user having to locate the file on their hard disk and the browser having to transmit it to the server. Once the user indicates that they have no more files to upload (the arc labelled finished), they are asked to *verify* the record they have just created by examining a mock-up of the page that would be displayed as a summary of this eprint record. The total elapsed time for depositing this eprint was six minutes and three seconds.

The variation in figure 3b shows a split-session deposit. Having filled in the metadata (3 minutes 26 seconds), the user, after step 2, elected to save the metadata entered up to that point. After a week (6 days, 22 hours) the same user elected to edit this eprint record and then went on to work on the metadata (just 1 minute and 24 seconds) before uploading the files as normal.

More complicated split-session sequences are commonly seen, for example in figure 3c there are instances where not only multiple editing sessions are recorded, but also multiple visits to the metadata page within a single editing session by the use of the ‘Previous’ button. By contrast, a degenerate tree (figure 3d) is seen where no file is uploaded to augment the metadata. Arcs that loop from the *metadata* node back on itself indicate that the user has missed out a piece of required information and has been returned to that form before being allowed to continue.



(a) Simple (b) Split session (c) Repeated metadata (d) No upload

Figure 3: Sample deposit visualisations

The ideal behaviour assumed by the software developers and Open Access campaigners is that shown in figure (a) or (b). The former would occur if the deposit is made after publication and all of the metadata and files are available. The latter would occur if a preprint were deposited, and the final publication data (journal volume, issue and page number) and article files were not available until some time later.

Findings

The displays allowed us to classify and list the most common interaction patterns as follows:

UNSUCCESSFUL ATTEMPTS: 66 attempted deposits (25% of the total sample of 260) did not terminate with a valid deposit, and in fact 20 of these ended with a confirmed deletion. The remainder were left in the depositing users’ workspace and were not visible in the public archive. This group may be an artefact of the data-gathering period cut-off time, and should perhaps be classified as ‘INCOMPLETE DEPOSITS’. (Only 2 of these records have gone on to be successfully deposited in the month after this study, and the remainder left in users’ workspaces.)

SUCCESSFUL DEPOSITS: Of the 194/260 (75%) records were successfully deposited; none were subsequently deleted.

OPEN-ACCESS DEPOSITS: 110 (57%) of the 194 successfully deposited records included at least one file upload, which means they provided open access to the article that the record describes. Only 19 of those full-text uploads occurred *after* the metadata had originally been deposited; the rest were uploaded at the same time as the metadata.

MULTI-SESSION DEPOSITS: 79 (41%) of the 194 successful deposits had at least one subsequent editing session to amend the previously entered record. (A subsequent session occurs when the deposit sequence is left and then re-entered through a specific request to edit the eprint record and not by continuing with one of the already-open deposit forms.) 22 of these multi-session deposits had their second and subsequent sessions after the deposit took place (*i.e.* after the 'Deposit' button was pressed.) By contrast, 54 had all their editing sessions before the deposit. The remaining 3 deposits had editing sessions both before and after the deposit session.

SINGLE-SHOT DEPOSITS: 89/194 deposits (46%) visited the metadata form only once. This may be considered the 'optimal minimalist behaviour', but only 38 of these single-shot sessions also uploaded a file with the metadata (in other words, this category may represent either efficiency and accuracy or laziness and minimal compliance!).

FORCED CORRECTIONS: 20 of the single session deposits were forced by the system to revisit the metadata page to fill in mandatory material that had been omitted.

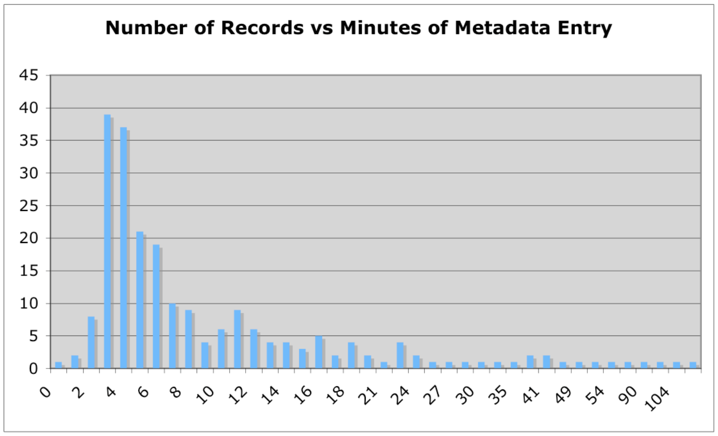
MULTIPLE EDITS: 6 of the single-session deposits had multiple voluntary visits to the metadata form (by use of the 'Previous' button) in order to revise data that had just been entered in the same session. Four of these appear to have been triggered by reading the 'Verify Record' page.

Metadata Timings

Entering metadata is a task that is resisted by researchers who have not yet done any self-archiving, and is thought to play a large part in the reasons that researchers resist using the repository. We analysed the combined timings for each of the metadata entry pages for each deposit, combining the various separate edit sessions and the various repeat visits to the metadata page (voluntary or involuntary). The time spent filling out the metadata page (ignoring the rest of the deposit sequence) was collated (a distribution of the timings can be seen in figure 4). *The median time for metadata entry is 5 minutes and 37 seconds per paper*. The average is 10 minutes 40 seconds owing to the long tail of the distribution.

The length of time taken to fill out the metadata correlates inversely with the number of records deposited by the user prior to this study (r = -0.25). That means deposit time shrinks as a user deposits more papers. An increase in efficiency with familiarity and experience is only to be expected; the extent to which other factors influence timing is not known.

The number of keystrokes was estimated for each record. Assuming that each required menu choice and button press counts for 2 keystrokes, while each text field contributes the number of characters it stores as its keystroke then the average and median number of keystrokes per record is 1500 and 970 respectively. (This is an overestimate because the paper title and abstract are frequently cut and paste from an online catalogue, or from the paper itself.)



Analysing the papers that were entered shows that, on average, each paper was co-authored by 3.33 individuals. Since only one of a paper's authors needs to self-archive it, an individual's nominal self-archiving time per paper will be, on average, scaled by a factor of 1/3.33. Investigation of wider self-archiving contexts [5,6] show that it is not just the researcher/author who undertakes the task of self-archiving themselves (63%) but librarians (21%) and students or assistants (12%) on their behalf. A researcher who writes one paper per month would accordingly find themselves (or their designees) spending an average of 12*10.66/3.33 i.e. about *39 minutes* per year in metadata entry tasks related to self-archiving. This figure would annually decrease as the authors' experience of self-archiving increases.

Conclusions

Self archiving is not a very time-consuming investment for the user (or designee) – about 10 minutes per paper, or just over a half-hour for a year's total research output. Growing evidence (cite REFS) of the consistent and substantial impact-enhancing effect of OA across all disciplines indicates that the return on this investment will be at least a 200% increase in citations and a knock-on increase in downloads and readings. Authors will be able to estimate for themselves how that increased research impact translates into the sorts of rewards they value (research progress, prestige, research funding, career advancement, salary). What is uncontestable is that it amounts to a small amount of time, well-spent, and yielding large returns.

The study described in this paper is limited to dealing with server logs, and so makes it impossible to distinguish between sessions that consist of prolonged attempts to enter metadata and those which consist of distractions and other tasks. Observed sessions and post-deposit interviews could be used to gain a more accurate understanding of the true figures for effort.

Further study is also necessary to discover the reason for the high percentage of undeposited material (the 66 "unsuccessful deposits" listed earlier).

The Open Access Keystroke Strategy requires that authors perform the keystrokes necessary for metadata entryand a full-text upload, even if the full text is not allowed to be distributed. According to this strategy the upload is *required* as a matter of institutional record-keeping policy for all articles (immediately upon acceptance), but the open-access key not required but strongly encouraged; also encouraged is preprint self-archiving. The N-1 Keystroke Policy is one that every institution can apply uncontroversially, and it is a strong antidote to the NIH delayed/embargoed back-access policy.

References

1. Barton, Jane, Currier, Sarah and Hey, Jessie M.N. (2003) Building quality assurance into metadata creation: an analysis based on the learning objects and e-Prints communities of practice. In 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice - Metadata Research and Applications, DCMI, 39-48. http://eprints.soton.ac.uk/20/
2. Harnad, S. & Brody, T. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals, D-Lib Magazine 10 (6) June
3. Hey, J. (2004) Metadata Issues for e-Prints: experiences from setting up an Institutional Repository. Presentation at ePrints UK Workshop, Ashmolean Museum Oxford, on 22 March 2004. http://tardis.eprints.org/papers/HeyJessie_eprintsuk_oxford22mar04.ppt
4. Hey, Jessie M.N. (2004) Targeting Academic Research: Southampton's Institutional Repository. In, Lewis, Jonathan (ed.) Proceedings of Online Information 2004, 30 Nov-2 Dec 2004, Learned Information Europe Ltd, 127-136. http://eprints.soton.ac.uk/13598/
5. Key Perspectives, Ltd. (2003) JISC/OSI Journal Authors Survey Report. http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf
6. Swan A. and Brown S. (2004) Authors and open access publishing. Learned Publishing, vol. 17, no. 3, pp. 219-224(6)