'Version control' of journal articles
by Sally Morris

The problem

Whereas discussion of Open Access journals may be maturing, and moving from 'religious' disputes to practical evaluation of alternative business models, the alternative route of Open Access – self-archiving – has been given impetus by high-level interest from the UK Science & Technology Committee and the National Institutes of Health and other funders. Until relatively recently, publishers were moving steadily towards permitting self-archiving in their agreements with authors; according to ROMEO and other studies, more than 80% of journals do permit authors to self-archive some version of their article. However, there is as yet no evidence that a significant majority of authors are actually doing so, other than in areas where there is a well established subject repository (such as ArXiv in high-energy physics and related areas). Nevertheless, changing policies (or even mandates) from funders and institutions could lead to a rapid change in author behaviour. When this is combined with the development of increasingly systematic ways of retrieving self-archived content, the threat to journal subscriptions could become acute. As Mary Waltham pointed out at the recent Journal Publishers' Forum, this will lead to an environment where 'near-substitute' versions of much published content will be freely available.

Some feel that the added value of the definitive, peer-reviewed, edited, formatted and fully functional version on the publisher's site will ensure that, even when he or she does not have free access through a pre-existing licence, it is still the first choice of the Google searcher; others are not so sure, and fear that a 'good enough' free substitute may gradually erode paid subscriptions and licences, potentially to the point where the parasite kills the host - these publishers are focusing their intention increasingly on other services which add value for the information-overloaded user. The work of CrossRef Search with Google is ensuring that, when both published and self-archived copies of the same article are found by a Google Scholar search, the published version appears at the top of the list. However, there is no established way of identifying the other versions, how much – if any – value has been added to them, and how they relate to each other.

This is also an issue for libraries, who share our wish to ensure that users access the definitive version, but who are also concerned about costs. Many libraries have gone to great expense and effort to implement link resolvers, attempting to point users to the 'appropriate copy' based on their institutional affiliations. But they are now confronted with the availability of different free online versions, and are wondering how to identify articles which have been placed in institutional or other repositories, or indeed self-archived elsewhere, and how to incorporate them into the OpenURL framework.

The difficulty is that these versions may differ in minor or even major respects from the published version; yet there is no way for the user to know this.

'Preprints', and indeed 'postprints', come in many variants, and we urgently need an agreed way to describe them, so that readers do know what they are getting. At the very least, there are the following versions:

- Privately circulated early draft (could be >1 iteration)
- Version presented at public event (again, could be >1)
- Pre-submission version(s)
- Version as submitted to journal x (may differ when re-submitted to journal y)
- Version amended after peer review (may go through >1 round of amendment)
- Version as accepted by journal x
- Accepted version, with substantive editing by journal editor and/or publisher (again, may be multiple iterations)
- Accepted version, with substantive editing and copy-editing - ready for publication
- Publication version (as above, formatted and paginated) – proof
- Publication version, corrected and passed for publication
- Published version, not on publisher's site (e.g. PDF), thus potentially lacking some functionality
- Published version (on publisher's site, with full functionality)
- Post-publication version with errata/ addenda (may be on publisher's site, with functionality), or elsewhere without it

This is an issue which was addressed in broad terms by an AAAS/STM working group as long ago as 1999/2000 (see http://miranda.ingentaselect.com/vl=2431145/cl=33/tt=885/ini=alpsp/nw=1/rpsv/~885/v13n4/s8/p251) but it has become increasingly pressing, and increasingly complex, since then. Work is needed both to agree nomenclature for the different pre- and post-publication versions of an article, and to establish metadata or other standards for their identification and linking. This work would probably best be carried out in collaboration with other publishing, library, user and standards
bodies.

What is proposed, therefore, is:

1. Analytical work to identify the different versions of a research article which can exist;
2. Proposed nomenclature to describe these;
3. Development of appropriate metadata to identify each variant version and its relationship to other versions, in particular the definitive, fully functional published version;
4. Establishment of practical systems for ensuring that these metadata are in fact applied (it seems unlikely that individual authors will consistently do so, but repository managers could and should).

Sally Morris, ALPSP
25/02/05