How to increase content in OA Repositories – What can be learnt from the special case of the research project PEER – Publishing and the Ecology of European Research

There are some findings and developments from the PEER project[1] on how to obtain more articles for institutional or subject-based OA repositories which should be of interest to COAR members. With the following remarks strategies for filling repositories are presented - for the very special case of PEER. The project is not aiming at gaining articles from the research community as a whole (from outside, as to say) but rather at being able to use a larger number of articles from a pre-defined set of journals and articles for its experimental purposes by employing internal strategies. Also, the example of PEER highlights some of the difficulties we have found along the way and how we have overcome them.

About PEER

Project design and aims

PEER (Publishing and the Ecology of European Research), supported by the EC eContent*plus* programme[2], has been designed to investigate the effects of the large-scale, systematic depositing of authors' final peer-reviewed manuscripts (so called Green Open Access or stage-two research output) on reader access, author visibility, and journal viability, as well as on the broader ecology of European research. The project is a collaborative effort between publishers, repositories and researchers and lasts from September 2008 to May 2012.

PEER Research

The project has commissioned and manages behavioural, usage and economic research which collectively are addressing such central issues as:

- How large-scale archiving may affect journal viability
- Whether it increases access
- How it will affect the broader ecology of European research
- Which factors influence the readiness to deposit in institutional and disciplinary repositories and what the associated costs might be

In sum, PEER Research[3] will provide objective input for evidence based policy making in the area of Green Open Access. To learn more about PEER research, please visit http://www.peerproject.eu/peer-research/.

Behavioural research surveys the attitudes of authors and readers for a scenario in which stage-two manuscripts are archived in repositories. *Economics research* examines the cost involved for publishers and repositories in making stage-two manuscripts available in open access. The *PEER Usage research* investigates actual usage of publishers' journals and repositories as evident from log files transmitted by publishers and repositories. Thus, it was necessary to wait until a critical mass of articles was available in repositories participating in the project before the usage research could commence.

[1] http://www.peerproject.eu/

[2] http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm

[3] http://www.peerproject.eu/peer-research/

PEER Observatory

PEER benefits from the active support of twelve STM publishers. To fill the observatory with content, participating publishers were collectively providing 241 journals for active participation of EU-authored manuscripts in the project. These journals cover the following four broad subject areas: life sciences, medicine, physical sciences and social sciences & humanities. On average, the participating journals were expected to have >40% EU content (corresponding authors based in the EU). The project centres on 'stage-two' articles, a stage-two article being defined as the author's final manuscript that has been accepted for publication by a journal and incorporates all the changes required by the peer-review process.

The participating journals were initially allocated to two deposit pathways for the project with each deposit route expected to provide up to 50% of articles:

a) Publisher submission – publishers are directly submitting both accepted manuscripts and associated metadata plus metadata associated with articles supposed to be submitted by authors to the PEER Depot and thus into the project.

b) Author submission – publishers invite authors to self-deposit their accepted manuscript. A special author submission interface has been created at the PEER Helpdesk[4] to guide authors through the submission process.

Submissions via both routes are deposited in the PEER Depot where they are processed. The PEER Depot which has been developed and is hosted by INRIA serves multiple functions within PEER. It matches author submitted content with publisher provided metadata, filters for EU research content which it holds for a specified embargo period prior to distributing to participating repositories. Additionally, it acts as a dark archive for PEER.

Each of the participating 241 journals has an embargo period set by the publisher, with factors such as subject area and individual journal economics being taken into consideration. Following successful processing, manuscripts are held by the PEER Depot for the duration of the embargo period agreed for each journal before distributing articles to participating repositories. With the exception of one social sciences 'subject repository' which is only receiving content from a set of disciplinary journals, each of the participating six PEER repositories hosts all valid PEER content, providing multiple mirror sites for the project.

Figure 1 depicts the PEER workflow with its two deposit routes.

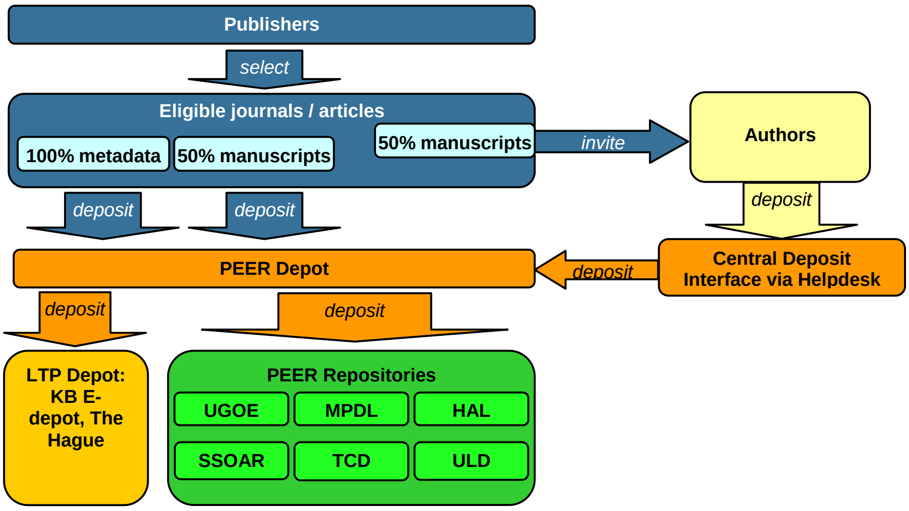

Fig. 1: PEER Workflow

[4] PEER Helpdesk http://peer.mpdl.mpg.de/helpdesk

Internal Content Reporting

Regular content reporting has been crucial for the project, especially while building a critical mass of content for the usage research. Biweekly content reports are used to monitor overall performance for the project, as well as being used to compare anticipated deposits per journal against actuals, with anomalies being investigated.

Implementation of the SWORD protocol

The SWORD[5] protocol (Simple Web-service Offering Repository Deposit) deposit mechanism offers a unified ingestion service and guarantees a robust transfer of manuscripts. In PEER, the SWORD protocol is used for article ingestion from the PEER Depot to repositories as it offers the option to transfer content to all partner repositories in a single simultaneous automated transfer process. Not only is this a new application in terms of the transfer of both metadata and full-text articles, it also results in a limited percentage of unknown errors in the transfer process and is thus a useful effort to achieve maximum content.

Further information on the SWORD protocol and its application in PEER and beyond can be obtained via the PEER Guidelines and the *Final report on the provision of usage data and manuscript deposit procedures for publishers and repository managers*[6].

This procedure for ingestion of material into repositories has applications beyond the project and may, in theory, be adopted by repositories to accept material directly from various sources, e.g. publishers, in the future.

In addition to the implementation of the SWORD protocol, PEER has developed a technical infrastructure that has applications beyond the project. This does not only apply for PEER partners but also for other interested parties. An essential ingredient of the PEER infrastructure are the *Guidelines for publishers and repository managers on deposit, assisted deposit and self-archiving* [7] which were particularly designed to fit the needs of participating publishers and repositories.

For additional tools and technologies which have potential applications outside of the finite duration of the PEER project see the *PEER Annual Report – Year 2*[8].

Acquiring Content – Issues

Within PEER it was crucial at a certain point in time to exceed a critical mass of embargo expired manuscripts required for the usage research in order to enhance the confidence levels of their research results.

Unfortunately, the project was experiencing some issues and therefore delays in accumulating enough content usable within the project due to reasons which could not be foreseen at the beginning of the project. Among other reasons, for some journals the percentage of EU-authored content turned out to be lower than anticipated. Also, a high amount of articles needed to be dismissed due to an article type unaccepted by the project.

In terms of publisher feeds one major issue was that it took longer than originally expected to set up a working infrastructure which also postponed usage research. The delay was due to the building of the Repository Task Force in PEER being less straightforward than expected. Besides dedicated repositories maintained by project partners (HAL, MPDL, UGOE) we were keen on

[5] http://swordapp.org/about

[6] Final report on the provision of usage data and manuscript deposit procedures for publishers and repository managers, 3 Nov 2009, http://www.peerproject.eu/reports/

[7] Guidelines for publishers and repository managers on deposit, assisted deposit and self-archiving, 28 May 2009, http://www.peerproject.eu/reports/

[8] PEER Annual Report – Year 2, Ch. 8, 30 Sept 2010, p. 5-6, http://www.peerproject.eu/reports/

finding additional repositories outside of the project to extend the geographic coverage in which content provided by the project would be made available. Regrettably, some repositories did not follow our invitation to join the task force due to various reasons.

All in all, delays within the project were due to innovation and change, technical and other challenges which could not have been foreseen at earlier stages of the project and were mainly due to its complexity.

Another major issue in PEER was the very low uptake of author self-deposit. Authors - of articles accepted for publication in a journal participating in PEER - were expected to collectively self-deposit up to 50% of the content into the project. Unfortunately, author deposit fell far short of expectations.

Increasing participating journal content in PEER

In support of building the volume of content available to PEER, with a focus on meeting the needs of the usage research team, substantial steps were taken to increase the volume of content available to the project from the participating journals.

After consultation with participating publishers, the following steps were implemented to further enhance the content position:

- 2 new publisher submission journals were added to the project, including back-content
- Embargo periods were reduced for selected journals (in some cases by up to 12 months).
- Additional back-content (manuscripts and metadata) was sourced from participating publishers

Throughout the lifetime of PEER it was necessary, however, to take even further steps for increasing the content available within the project. Much attention had already been spent on author communications and author support within PEER. Due to the short timescale remaining in the project, and the long lead times likely to be needed to change author behaviour, it was decided instead to focus on improving the content available for research within PEER. It was therefore agreed that, where technically feasible, a number of 'author deposit' journals would transfer to the 'publisher deposit' pathway.

- 48 'author deposit' journals were transferred to the 'publisher deposit' route
- Back-content for 20 'author deposit' journals from one publisher were successfully processed by the PEER Depot as if they were following the 'publisher deposit' route, while retaining the live feeds within the 'author deposit' pathway.
- Manuscript types accepted in the project were reviewed and increased.

Also investigated was the inclusion of EU co-author manuscripts. However, the programming required was more complex than originally thought and would have taken well after the deadline set for acquiring the critical mass of content for the usage research. Therefore, efforts were concentrated on other options.

Extraction of Metadata from PDFs

Within PEER, articles can only be used, when attached with metadata. Where a full set of metadata was missing (crucial: DOI and publication date) PEER has benefitted particularly from the development of *GeneRation Of Bibliographic Data (Grobid)*. Grobid is a text-mining tool enabling the automatic metadata extraction of bibliographic metadata from PDF[9]. INRIA has worked with a participating publisher to enrich the mandatory metadata provided via extraction of additional metadata from PDFs provided by this publisher. This task is complementary to the

[9] More information on Grobid can be obtained at http://grobid.no-ip.org/

publisher's data transformation task and has only been performed when crucial metadata were missing.

This was a prospective action to extract information automatically from a PDF file. To fulfil this task GROBID environment was used and trained to match various title page styles in scholarly papers. Particularly good results were obtained allowing the automatic ingest of PDFs within the PEER Depot for all documents provided by some publishers. The process has been used to extract metadata for over 900 publisher submitted manuscripts and over 1,000 metadata files were extracted for possible matching with author deposited manuscripts.

The development of the Grobid process for metadata extraction from PDFs is an exciting development for PEER which is likely to have many future applications after the project's conclusion. This adds to the growing number of examples of technical solutions created within PEER to overcome the challenges encountered with a large scale green open access initiative.

Self-Archiving

The project made extensive efforts to engage with authors and to facilitate author deposits via the author submission interface. Also, the PEER Helpdesk was intended to serve as the central point of communication. For reasons of data privacy, no direct communication was intended between project members and eligible authors. Thus, communication texts were sent by publishers at acceptance of the article inviting authors to participate in PEER and to deposit their articles. Despite these efforts, the percentage of responses from authors who actually self-deposit within PEER remain low (*figures as of April 2012*):

Invitations to authors to self-deposit: >11,800
Author Deposits: 170
% authors depositing: <2%

The Behavioural Research conducted within PEER lead to striking results regarding authors' deposit behaviour. Motivations and barriers for authors in terms of self-depositing are disclosed in the PEER Behavioural Research Team's Baseline[10] and Final[11] reports.

Barriers to author self deposit in an OA repository

The ethos underlying open access repositories was appealing to most participants questioned within the framework of our Behavioural Research, but lack of knowledge was explicitly articulated as a barrier. This applies to both lack of knowledge of appropriate repositories to deposit in and of the deposit process itself. Some studies furthermore suggest that self-archiving behaviour is primarily affected by disciplinary norms. Also, researchers are concerned about what to expect from repositories in terms of prestige and quality of content. What is more, authors fear to infringe copyright which is perhaps the most crucial barrier to self-deposit[12].

Tedious deposit procedure

The deposit process itself is often cited as a barrier towards self-deposit by authors. Maybe due to a lack of knowledge, depositing articles is regarded as a tedious, extremely time-consuming and somehow discouraging procedure by researchers, especially when it comes to sorting out copyright issues. Time pressure, technical difficulties and learning new processes (which can range from file upload to metadata input) are all part of the deposit process and may all contribute to authors' concerns about depositing material in an open access repository.

[10] PEER Behavioural Research - Baseline report, 1 Feb 2010, http://www.peerproject.eu/reports/
[11] PEER Behavioural Research - Final Report,06 Sept 2011, http://www.peerproject.eu/reports/
[12] Cf. for this chapter PEER Behavioural Research - Baseline report, ch. 4.1.1 & 4.1.2.

Article quality

Researchers are apparently uncertain about depositing in a repository and about what to expect from articles they will find there in terms of quality, be it the reputation of a repository itself in the case of authors, or the quality of the articles deposited in the case of readers. The main concern seems to be that if repository content was fragmented in terms of quality, this will have a negative influence on the overall prestige of the repository and thus reduce the visibility of authors contributing to that content.

Also, researchers often feel that open access material is of lower quality than content published physically and liable to pay, more precisely open access works are often associated with non-peer-reviewed materials[13].

Copyright issues & article version

Copyright and publishers' self-archiving policies are perceived as a major barrier to self-archiving. Generally, authors do not know whether they have permission to upload a copy of their article onto a repository or website, nor do they know which version (pre-print, author's final copy or publisher's copy) they can deposit.

Fig. 2 summarizes barriers to repository deposit associated with researchers' assessments.

Figure 2.9 Barriers to repository deposit

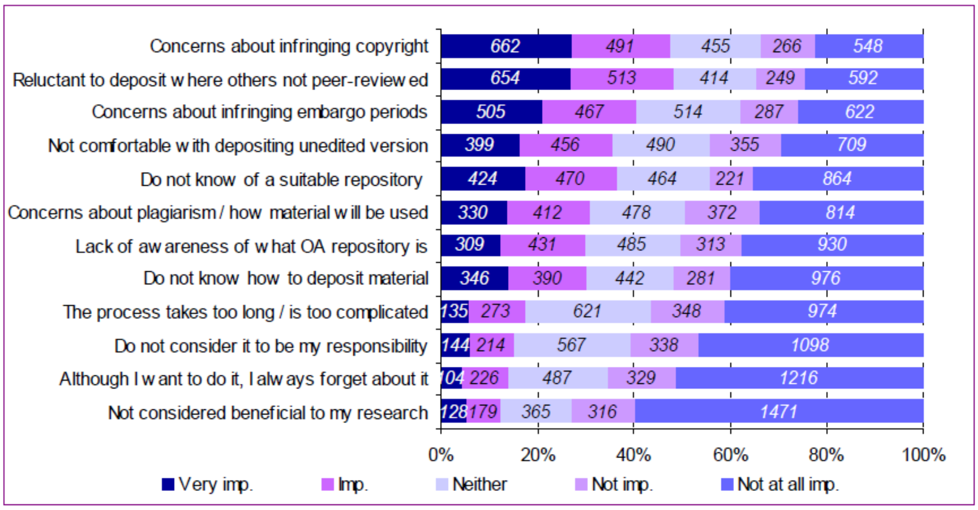

Fig. 2: Barriers to repository deposit (PEER Behavioural Research: Baseline report, p. 35)

Incentives to author self-deposit in OA repositories

Motivations to self-deposit may be serving the public or the greater good, but may beyond that be in the authors' own interest. First of all, repositories grant free and unlimited access to all, but they also guarantee widespread availability, a higher speed of dissemination and increased citations for the authors' outputs. Not without reason are deposits in a repository regarded as an important factor in terms of career advancement by some researchers. Factors encouraging OA repository deposits are shown in figure 3 and are also set out in the Final report of the PEER Behavioural Research, Ch. 4.1.4.

[13] Cf. PEER Behavioural Research - Baseline report, p. 16.

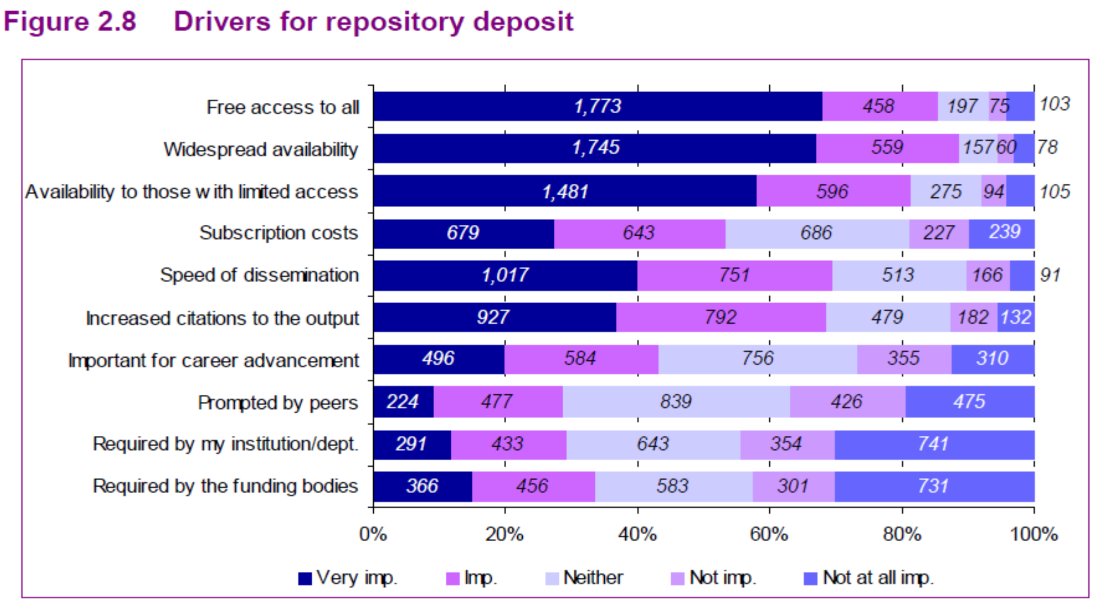

Fig. 3: Drivers for repository deposit (PEER Behavioural Research: Baseline report, p. 34)

Making researchers better aware of the obvious advantages of repository depositing also holds potential to boost a repository's growth. Furthermore, it seems to be important to 'make it easy' for authors to deposit, perhaps by entrusting library or other staff with this task.

Moreover, it would appear that publisher-mediated deposit and clear guidance from publishers as to their attitude towards deposit of published articles and the version that may be made available in a repository, would be of considerable help in overcoming author concerns.

Mandatory author deposits might also be an encouraging factor, since in recent years several universities as employers of academic staff and funding agencies in most disciplines have developed open access policies or at least position statements on open access. Furthermore, where funding agencies stipulate mandatory deposits they also encourage grant applicants to apply for the necessary funds to make articles publicly available via open access.

There is a large community advocating mandatory author deposits, with mandates being demanded by funders as well as by institutions, as a decisive way to obtain content for repositories. In PEER's Behavioural Research, however, it has been observed that funder or institutional mandates were considered relatively unimportant as drivers for repository deposit by survey respondents[14].

It is obvious that appropriate action from publishers is an efficient method to populate repositories. In the PEER experiment, publishers' deposits provided the vast majority of articles for participating repositories compared to authors' self-deposits which only contributed a negligible number of articles. The number of author deposits in PEER is so low that it becomes significant for our assessment of self-archiving as the only solution to fill repositories.

Concluding remarks

The PEER project has been designed to investigate the effects of the large-scale deposit of stage-two articles in repositories. To this end, an observatory has been set up to exercise a deposit experiment. During the project it was necessary to increase the volume of content dramatically

[14] PEER Behavioural Research - Baseline report, 1 Feb 2010, p. 70.

at a certain point in time which resulted in certain additional efforts and considerations within the project.

The PEER experience shows us that it is essential to convey positive messages and incentives for authors to self-archive their manuscripts: *What is in it for them if they self-deposit?*

In addition, it is necessary to find creative strategies by using new technologies: *What tools can be used to facilitate the deposit process?*

Besides author deposit an additional approach is needed to gain a meaningful amount of content for repositories. In short, self-deposit is perhaps a good starting point to fill repositories and desirable for many reasons, but additional strategies like publisher-mediated deposit might be recommendable, not least to obtain a reliable version of record, reliable metadata and a constant and regular flow of articles.

The PEER project members hope that COAR's members and partners will benefit from the strategies outlined in this experience report.

April 2012, Dr Barbara Bayer-Schur, SUB Goettingen

If you are interested in further information on PEER, please do not hesitate to contact me at [mailto:bayer-schur@sub.uni-goettingen.de]

Literature

Reports:

– Final report on the provision of usage data and manuscript deposit procedures for publishers and repository managers, 3 Nov 2009, http://www.peerproject.eu/reports/

– Guidelines for publishers and repository managers on deposit, assisted deposit and self-archiving, 28 May 2009, http://www.peerproject.eu/reports/

– PEER Annual Report – Year 2, Ch. 8, 30 Sept 2010, http://www.peerproject.eu/reports/

– PEER Behavioural Research - Baseline report, 1 Feb 2010, http://www.peerproject.eu/reports/

– PEER Behavioural Research - Final Report, 06 Sept 2011, http://www.peerproject.eu/reports/

Online resources:

– PEER project website http://www.peerproject.eu/

– PEER Research webpage http://www.peerproject.eu/peer-research/

– EC *E*Content*plus* programme http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm

– SWORD protocol http://swordapp.org/about

– GROBID http://grobid.no-ip.org/