Microsoft Academic integration within EPPI-Reviewer

What the integration involves, and how to use it. DRAFT v.3

Introduction

The EPPI-Reviewer database system now contains a copy of the Microsoft Academic[1] dataset. Similar to Google Scholar, Microsoft Academic aims to be a comprehensive repository of the World's research and currently contains more than 250 million records. Unlike Google Scholar, it is possible to obtain a copy of the dataset; a copy that is updated every couple of weeks.

The availability of a comprehensive, regularly updated, repository of research could be extremely useful for systematic reviewers in three ways. First, given its comprehensive nature, it probably contains all, or nearly all, the studies that are likely to be relevant for almost any systematic review, thus reducing the number of sources that reviewers need to search. Boolean searches can be carried out to *identify studies that might be relevant in new reviews*. Second, when some relevant studies have already been located, connections in the graph of publications can be used to locate other studies that are 'close' to the studies already known to be of interest. This feature can be used to *bring a review up to date* and also for *citation chasing* – where reviewers check bibliographies of known records for other eligible studies. Finally, as the dataset is updated regularly, it can be used as a way of *keeping existing systematic reviews up to date*, supporting 'living' review workflows. The challenge facing both use scenarios – and not an inconsiderable one! – is how to distinguish the records of interest from the vast majority of other (irrelevant) records. We are therefore releasing a set of tools within EPPI-Reviewer that will enable reviewers to access the Microsoft Academic dataset and evaluate its utility in their specific reviews. This is to be considered "work in progress", with the aim being to support research and development, rather than necessarily being ready for use in all reviews.

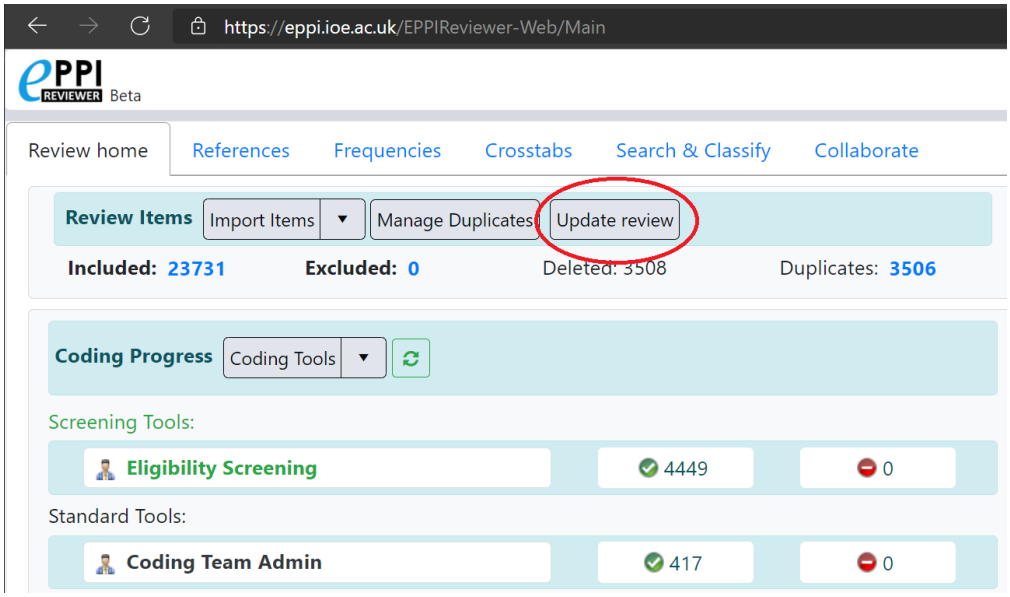
This document outlines how the new features in EPPI-Reviewer can be used in all three scenarios. This should be considered 'Beta' software, and we welcome feedback on how it performs in different situations, and how it might be improved.

[1] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MA) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 243-246. DOI=http://dx.doi.org/10.1145/2740908.2742839

Getting started

In order to begin to use these features in EPPI-Reviewer click the 'Update review' button which is on the home page (Figure 1).

Figure 1: location of the 'Update review' button



Keeping a review up to date ('living systematic reviews and maps')

One of the most powerful new features available is the ability to *keep a review up to date* using the Microsoft Academic dataset.

Every two weeks a new copy of the database arrives, with up to around a million new records. If your review is 'subscribed' to the auto-updating service, a machine learning model will 'learn' the scope of your review, and automatically suggest new records that might be relevant. You subscribe to this service on the 'keep review up to date' tab, selecting whether the machine learning model should analyse ALL the items in your review, or only those with a specific code. Most reviews will need to select only items with a specific code, as your review probably contains both items that are, and are not, relevant to your review. **It is important to ensure that the machine 'learns' only from those items that are actually relevant to your review.**

Each time the service runs, a new row will appear in the list entitled 'Items found at each task execution' with the most recent listed at the top. You can then click on the 'refine / import' link to decide what to do with the new items that the machine learning service has identified.

As Figure 2 shows, you use the options available on the 'refine / import' page to determine which items you want to bring into your review. The first graph that appears here are the results from the 'auto-update' algorithm. This is the algorithm that examines all the newly arrived records and determines whether any of them might be relevant for your review. It is a very sensitive algorithm, and so will probably produce many more candidate records than you want to examine; its purpose is to find all *potentially relevant* records and leave fine-tuning to subsequent steps.

These subsequent steps involve one or more of the following:

- importing the top n records according to the above algorithm (where 'n' is a number determined by each review independently depending on screening capacity and previous evaluation of the use of the auto update scores
- using one of EPPI-Reviewer's study type classifiers either to rank records according to their relevance, or to use a cut-off threshold, below which records will not be imported
- using a 'user' classifier, which is a classifier built using review data

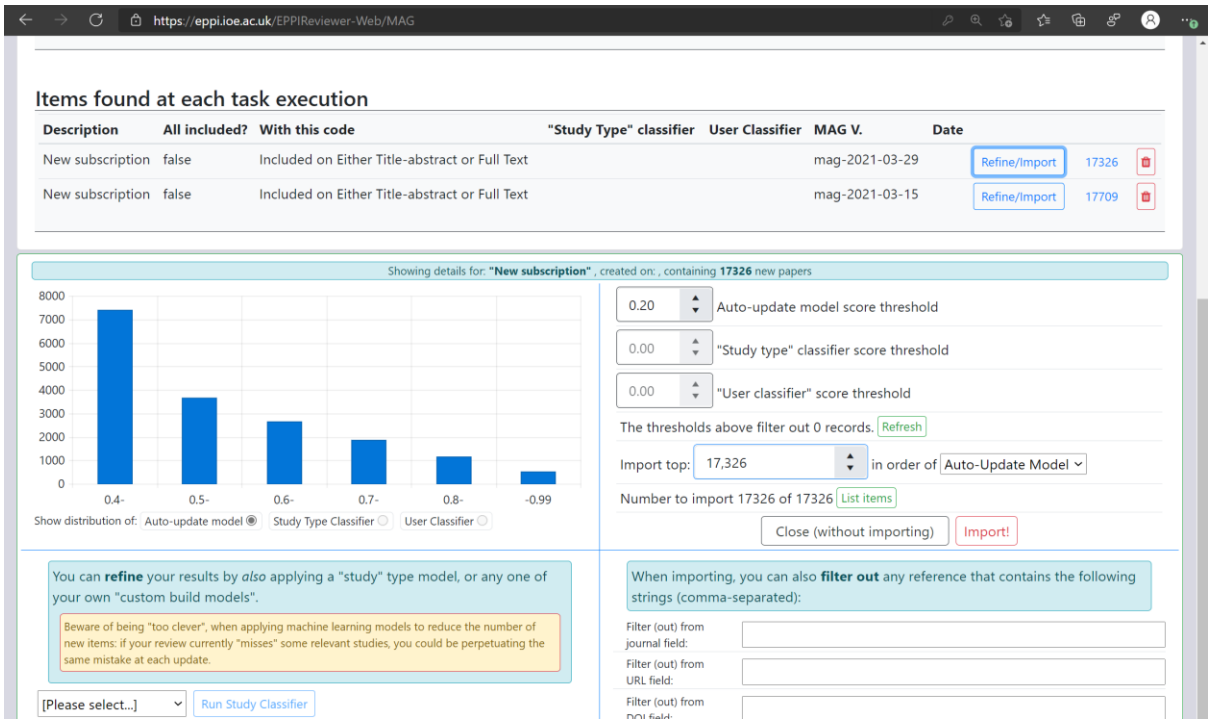
For example, a workflow that aims to identify randomized controlled trials that evaluate smoking cessation interventions might classify all records using the RCT classifier, and also all records using a user classifier that was built using the included and excluded studies from the original review. It would set a cut-off on the RCT classifier and exclude all records that fall below a given threshold (e.g. 0.1) and then import the top n items ranked according to the user classifier. This would maximise the chances of identifying relevant records by filtering out all records that were clearly the wrong study type, and then importing those that were most likely to be relevant to the review according to a classifier built using its data. The review authors might then use priority screening to screen the resulting records (though would aim to keep them to a reasonably small number, as they could be importing similar numbers of records every two weeks).

If you are not looking for RCTs (or systematic reviews or economic evaluations), you have two choices. You can either import the very highest scoring items in the list into your review, or you can run a 'user classifier' that can be built using your existing include / exclude decisions. We

recommend that you experiment with ranking new records by auto update and user classifier scores and examining which thresholds might work best for your particular use case. It's likely that a combination of setting a threshold (e.g. 0.5) on the auto update model score and then ranking the records according to your user classifier will give you the best results.

Please bear in mind that this is a very new feature, and methods and tools are still developing. We are therefore being far too over-inclusive in the number of records retrieved at the moment. In the example shown in Figure 2, there are about 17,000 new candidate records arriving in that review. Bearing in mind that these kinds of numbers arrive every two weeks, we are not suggesting that they all need to be looked at. The aim of the over-inclusivity is to enable experimentation using the study type and user classifiers, so please don't be put off by the dauntingly large numbers of records being identified! Our aim is to support methods development here, and it is unlikely that more than a handful of relevant records are published each week: the handful of records is likely to be in the set retrieved, and the ranking algorithms aim to bring them to the top of the list.

Figure 2: deciding which items to import

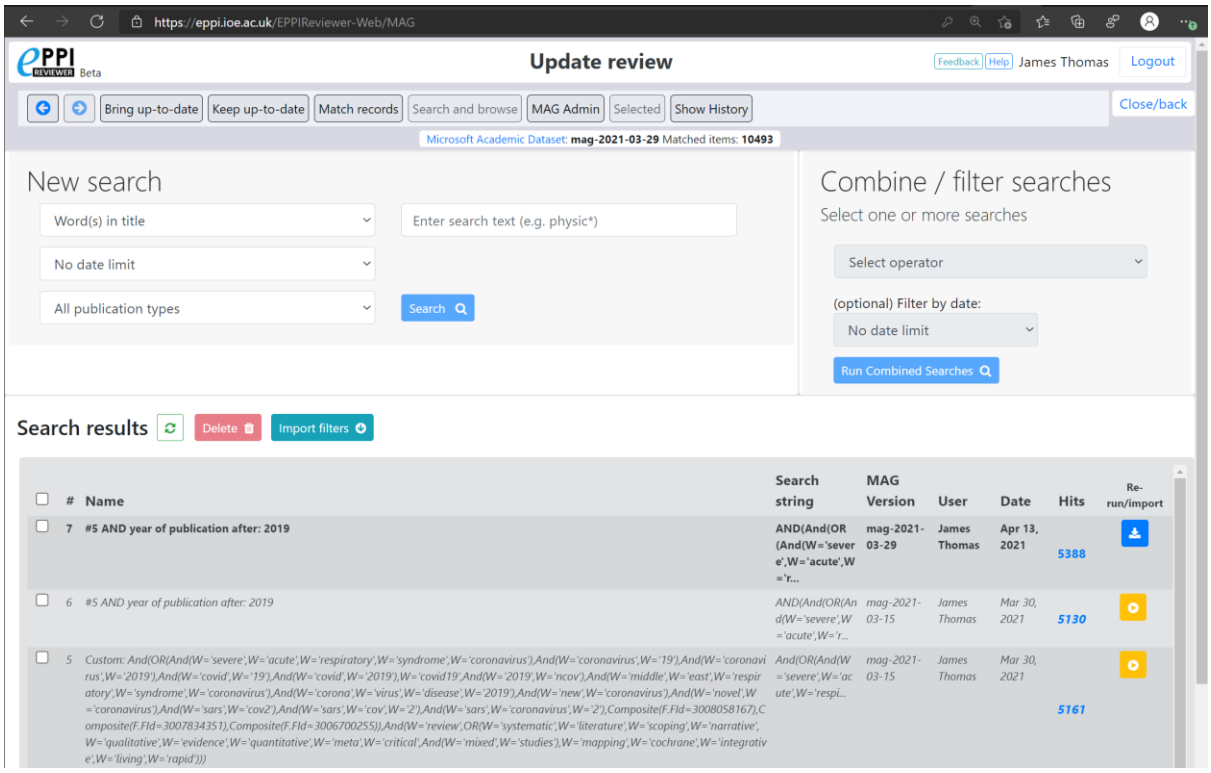


Boolean searches of Microsoft Academic

It is possible to carry out standard Boolean searches of Microsoft Academic (Figure 3) using:

- words that appear in titles and abstracts
- the names of authors and journals
- automatically generated 'topics'

Figure 3: Boolean search



It is also possible to filter search results according to publication date and year ranges. Search history is stored in a list and searches can be combined using AND and OR. While this is a very powerful search engine across the whole of science, there are some limitations:

- no 'NOT' operator
- no wildcards
- no phrase searching (e.g. you can search for "systematic" AND "review" in the title field, but not "systematic review"

It is possible to import up to 20,000 records at a time from the search results. Importantly, no duplicate items will be imported, meaning the same search can be run multiple times over time and new results imported.

Bringing a review up to date and performing 'citation chasing'

The third scenario to consider is when we want to bring an existing review up to date and performing 'citation chasing'. In this situation we already have lots of information about the review, including the studies that have been included and excluded. We can use this information to help us find new and related studies efficiently. *Before we are able to use these advanced features though,*

as many records in the review as possible need to be 'matched' to their equivalent in Microsoft Academic. Please see the section below on ***matching records to Microsoft Academic****.*

Assuming you have the records included in your review matched to their Microsoft Academic equivalents, you can use the features described in this section to bring your review up to date.

First of all, you can use the *automatic Boolean search generator*[2] to create a Boolean search for you, based on the studies already included in your review. This function analyses the studies you have previously included and attempts to create a Boolean search that is capable of identifying as many of the records you have included as possible, without including vast numbers of irrelevant records. The algorithm analyses both the titles and abstracts of your included studies as well as the 'graph' around your studies: the networks of citations and recommended studies. After running the Boolean search, and importing the new records into your review, you can use the 'priority screening' function to rank the new records automatically in terms of which is most relevant to your review.

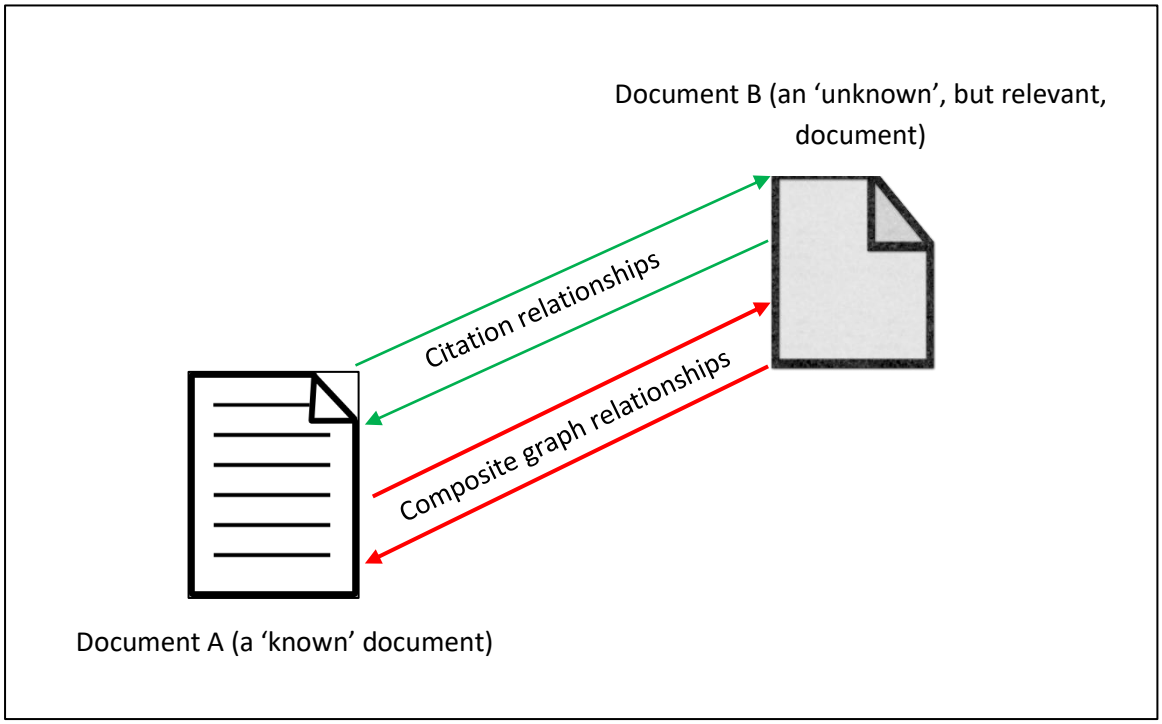
The second method involves using the 'graph', or 'network', of publications in Microsoft Academic, all of which are related to one another through their citations, semantics, authors, place of publication, and institutional affiliations. If we start with one or more 'known' publications, we can follow their relationships to find other, similar publications. This is sometimes known as 'snowball' searching, and while it can be efficient, systematic reviewers tend not to rely completely on these approaches because of potential bias. (e.g. when a relevant document is not cited by other relevant documents) We are currently still evaluating the best approach to take in relation to 'graph' searching, but present six possible options, depicted in Figure 4.

Figure 4 outlines two types of relationships between documents: 'citation' and 'composite graph'. The use of citation relationships has been discussed many times over the years, as it can be an efficient way of finding relevant studies, based on studies that are already 'known'. It is possible to follow citation relationships in two directions: first, the papers listed in the bibliographies of 'known' records; and second, papers which cite the 'known' records. In addition, Microsoft has analysed the large number of different ways in which documents can be related to one another, and has created a composite of these relationships known as the 'related documents' feature. Any document can have a maximum of 20 other related documents and, again, these relationships can be followed in either direction.

While there are concerns about using citation networks as being a single method of searching, they are a valuable and valid part of a systematic search strategy. As long as your records are 'matched' against their Microsoft Academic equivalents, it's easy to conduct bi-directional citation chasing within EPPI-Reviewer (a task that takes much longer if done manually!).

[2] This algorithm is not yet in the user interface, but will be ready soon.

Figure 4: finding related documents using a graph search



The Microsoft Academic Browser

Lists of records from the Microsoft Academic can be viewed in the 'Microsoft Academic Browser' – a built-in way of browsing the Microsoft Academic dataset within EPPI-Reviewer. This browser enables users to view all the information held about specific papers, browse the dataset by topic, and also to select records into a list which can later be listed and/or imported into the review.

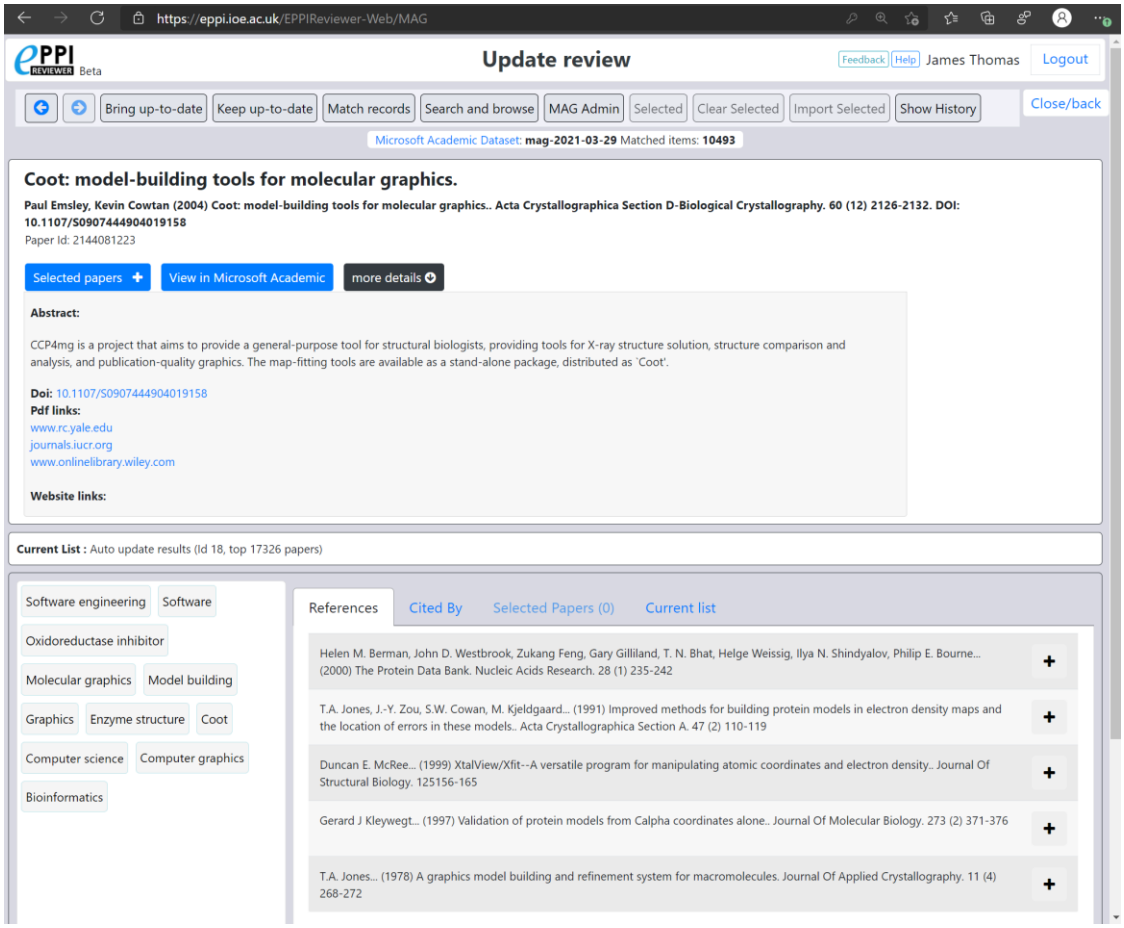
Selecting records

First, it is possible to 'select' records into a list that can then be imported or examined later. The number of records in the list is shown at the top of the screen. Note that this list is not saved to the database, and so is reset if you close your browser window. Using the links at the top of the screen you can clear the list of selected items or list them. Records that are already in your review will be listed among the search results, but can't be added to the list of selected items.

Browsing by topic

Second, you can click the 'topics' on the left hand side of the screen to examine the records that have been automatically classified as belonging to that topic by the Microsoft Academic algorithm (Figure 5). You can also search for topics in the Boolean search. Topics have a hierarchy, so you are able to navigate 'up' and 'down' the hierarchy. The topics themselves have been generated automatically too, so they may change from time to time, as the algorithm is re-run on new data, and the models are updated.

Figure 5: detailed information about a specific paper



Information about a specific paper

Figure 5 shows some of the detailed information about a specific paper contained in the database. Each document might be found in more than one place on the web (e.g. on a journal's website and in an institutional repository), and this page will list all the URLs for that paper. It will also present lists of the papers included in that paper's bibliography, the papers that cite it, and those that are related to it[3], according to the composite graph relationship algorithm. The topics associated with a specific paper, and whether or it is already in your review are also stated.

Matching records to Microsoft Academic

Many of the above functions depend on review records being matched into their equivalent records in Microsoft Academic. This is so that the metadata held by Microsoft Academic about the review papers can be utilised. Records can be matched automatically, using the 'match records' page. The majority of journal records should be matched automatically, as long as they contain sufficient data. The automatic matching function uses the same algorithm as the deduplication algorithm, and accepts any match as a 'true' match if it scores above 0.8. Matches between 0.5 and 0.8 are referred to users for manual checking (Figure 6).

[3] This feature will be implemented when the data are available in the index

Sometimes, the record can be found on the Microsoft Academic website, but has not been listed even as a possible match by the matching algorithm. In these situations it is possible to enter the appropriate Microsoft Academic ID and add the match manually (Figure 6).

Figure 6: checking and adding matches between review records and Microsoft Academic