Microsoft Academic Graph Browser in EPPI-Reviewer – User Guide (v1.0)

What MAG Browser v1.0 can do and how to use it

Introduction

EPPI-Reviewer online systematic review software now contains a copy of the full *Microsoft Academic*[1] dataset. *Microsoft Academic* is a large Open Access (OA) repository containing >228 million OA bibliographic records of research articles from across science, connected in a large network graph of conceptual and citation relationships – the *Microsoft Academic Graph* (*MAG*). Microsoft makes the *MAG* dataset available for 3rd party use under a creative commons license, updated with records of new research articles every 10 days. As well as standard bibliographic information, *MAG* records include hyperlinks (URLs) to corresponding full-text study reports identified on the internet.

If the vast majority of study reports needed for systematic reviews are accessible and easily discoverable in *MAG*, then the focus of electronic search methods could shift away from conventional Boolean searches of multiple electronic literature databases, towards the efficient data mining of this single, large OA repository. If realised, such a shift could also achieve large efficiency gains in systematic review production and updating systems, by establishing a semi-automated, prospective surveillance system for 'new' eligible study reports that will help ensure reviews – including 'living systematic reviews'[2] – and their findings can more easily be kept up-to-date. The key challenge in using *MAG* to support efficient study identification in systematic review and other use scenarios – is how to distinguish the typically small set of 'target' *MAG* records of interest from the vast majority of other (irrelevant) records in the full dataset.

We have therefore released a suite of new tools within EPPI-Reviewer that enables users to access the full *Microsoft Academic Graph* and evaluate its performance for efficient study identification in their own systematic reviews and other use scenarios. These tools are deployed in a new user interface (UI) in EPPI-Reviewer called ***MAG Browser v1.0***. This UI includes ***MA Graph Search v1.0*** – a new search tool (containing a graph analytic prioritisation algorithm) that harnesses relationships, structure and other features of *MAG Network Graphs* in order to identify 'target' *MAG* records with acceptably high levels of recall and precision, compared with conventional methods.

MAG Browser v1.0 – incorporating *MA Graph Search v1.0* – are still very much "works in progress". That is, we are making these tools available with the primary aim of supporting further their research and development in a range of potential use scenarios – and they cannot yet be considered

[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

[2] For further information about 'living systematic review' methods, tools and workflows, please visit: https://community.cochrane.org/review-production/production-resources/living-systematic-reviews.

tools ready for general use in all systematic reviews. *MAG Browser v1.0* should therefore be treated as 'Beta' software; and we welcome feedback on how it performs in different use cases, and/or how it might be improved.

Please send all feedback on *MAG Browser v1.0* and *MA Graph Search v1.0*, via e-mail to EPPISupport@ucl.ac.uk – including 'MAG' in the e-mail subject line.

This document outlines what *MAG Browser v1.0* can do and how its new tools and features – including *MA Graph Search v1.0* – can be used in EPPI-Reviewer to support study identification in two specific, selected use scenarios: 1) 'Finding additional ('new') studies'; and 2) 'Continuous updating of a living systematic review'[3].

In order to use and evaluate *MAG Browser v1.0*, you will first need to:

- Login to EPPI-Reviewer at https://eppi.ioe.ac.uk/cms/er4;
- Create a new review or open an existing review; and
- Import your 'known' bibliographic title-abstract records into the review (if not already done) and assign them to a new code (if not already done). The set of 'known' records could, for example, comprise all of the included study reports in a published systematic review.

If you are unsure how to complete these preparatory steps, please consult the EPPI-Reviewer User Manual – available from: https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2933; and our 'instructional videos' on YouTube – available from https://eppi.ioe.ac.uk/cms/er4 (please also visit the latter webpage if you do not yet have an EPPI-Reviewer user account[4]).

[3] These two use scenarios were selected for the sole purpose of introducing *MAG Browser v1.0* – that is, *MAG Browser* can also be used in various other 'study identification' use scenarios beyond systematic reviews (and beyond the scope of this document).

[4] EPPI-Reviewer is available 'free at the point of use' to Cochrane authors and Campbell authors (via an *Archie* authentication) for use in the production of both new and updated Cochrane or Campbell reviews. For further information, please visit: https://community.cochrane.org/help/tools-and-software/eppi-reviewer.

Use Scenario 1: Finding additional ('new') study reports

Our first use scenario broadly focuses on using *MAG* to find additional ('new' or 'unknown') eligible study reports – for example, when updating a systematic review. As described above, *MAG* records are connected to one another in a very large network graph. Therefore, if we 'start' with one or more 'known' *MAG* records (i.e. 'nodes' in the graph), then we can follow the relationship 'links' (i.e. directed 'edges' in the graph) to find other, connected publications. This general approach of generating and searching *MA Network Graphs* from 'known' study reports (*MAG* records) to identify 'new' (or 'unknown') study reports (*MAG* records) is conceptually similar to 'snowball searching'. However, it also overcomes known limitations of the 'snowballing' method, because it does not rely exclusively on (forwards and backwards) citation relationships (see below). We are currently still evaluating the best option(s) for generating and searching *MA Network Graphs*, among the seven initial options depicted in Figure 1.

Figure 1: Finding additional ('new' or 'unknown') documents using a 'graph search' approach

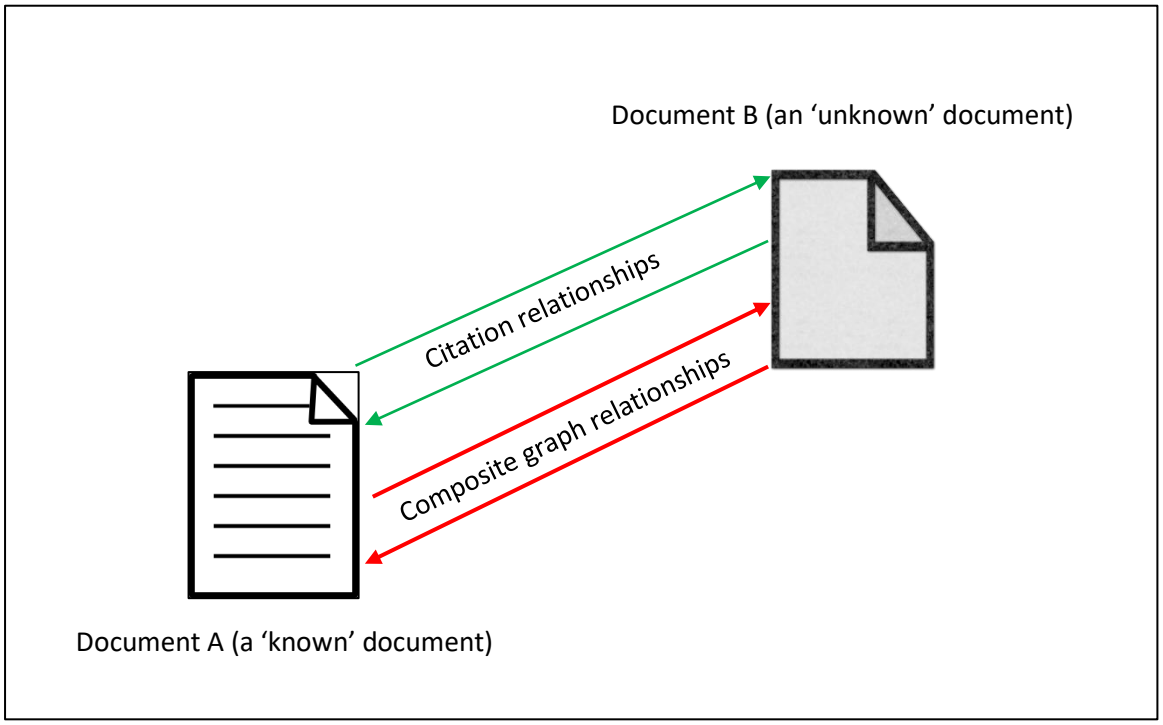


Figure 1 outlines two types of network graph relationships between one or more 'known' (Document A) and one or more 'unknown' (Document B) documents in MAG (i.e. *MAG* records): 'Citation' relationships and 'Related Publications' (aka 'Composite Graph') relationships.

It is possible to follow 'Citation' relationships in two directions (i.e. they are 'bi-directional'): first, the documents listed in the bibliographies of 'known' records; and second, documents 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 within *MAG*; and has created a 'composite' of these relationships known as the 'Related Publications' feature. Any document can have a maximum of twenty other related documents; and again, these relationships can be followed in either direction (i.e. they are 'bi-directional').

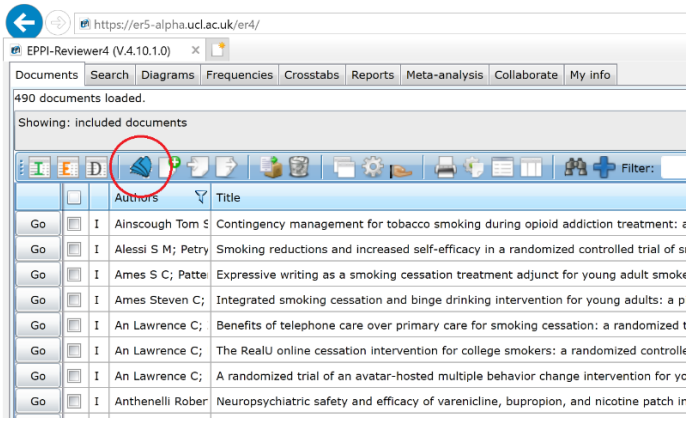
MAG Browser v1.0 therefore currently offers seven variant options for 'finding additional ('new' or 'unknown') study reports in using *MAG Network Graph* relationships:

1. 'Unknown' documents 'recommended' in the 'Related Publications' lists of 'known' items ('Recommended by');
2. 'Unknown' documents that 'recommend' 'known' items in their 'Related Publications' lists ('That recommend');
3. Bi-directional recommendation relationships ('Recommendations') – this option combines '1' OR '2';
4. 'Unknown' documents in the bibliographies of 'known' items ('Bibliography');
5. 'Unknown' documents that cite 'known' items ('Cited by');
6. Bi-directional citation relationships ('Citations') – this option combines '4' OR '5'; and
7. Bi-directional recommendation or citation relationships ('Recommendations or citations') – this option combines '3' OR '6'.

In each case, *MAG Browser* can be used to generate a *MA Network Graph*, from the initial set of 'known' study reports based on the selected 'graph' relationship(s); and then identifies and retrieves all the 'new' ('unknown') MAG records that are present in the graph.

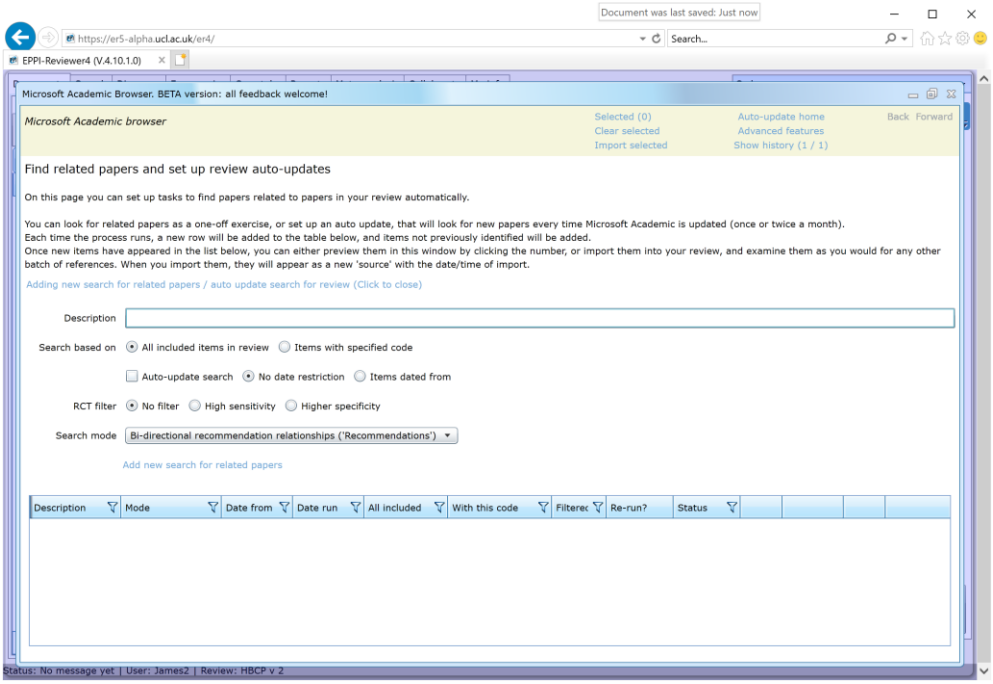
In order to begin to use these features and options in EPPI-Reviewer, click the *Microsoft Academic* icon on the home page of your review (Figure 2).

Figure 2: Location of the MAG Browser launch icon



You will arrive at the page which enables you to search *MAG* for 'new' study reports and set up 'auto-updates' (more on this later). To search for 'new' study reports using the seven options for generating *MA Network Graphs* described above, click 'Add new search for related papers…' and the screen shown in Figure 3 will appear.

Figure 3: Adding a search for related documents



First, you can give your search a name, and then configure it using the following options:

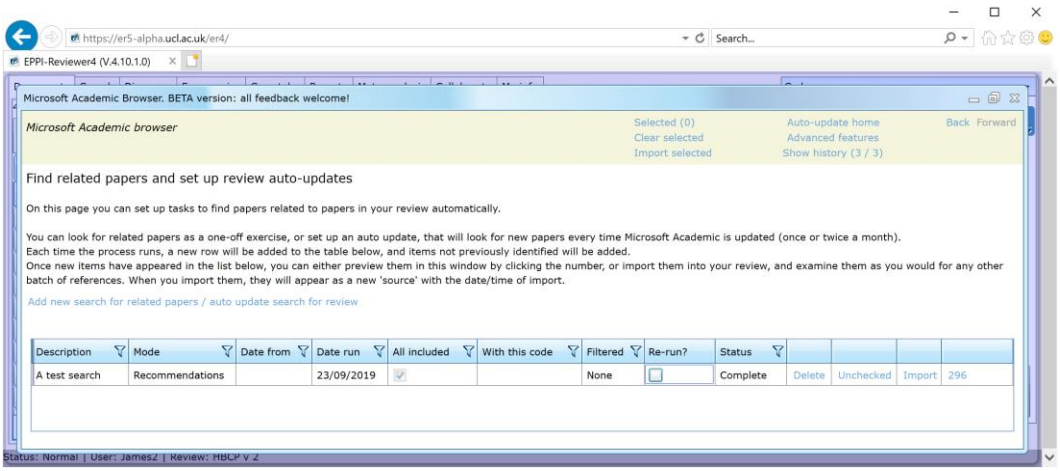
- It can be based on all the included records in your review, or only those with a specific code. This refers to the pool of 'known' documents shown as 'Document A' in Figure 2. The search will 'start' from these documents and identify records which are connected to them.
- You can decide to 'auto-update' your search. This supports 'living systematic review' workflows by re-running the search each time a new update of the full *Microsoft Academic* dataset becomes available (approx. every 10 days).
- You can also set a date filter on the search. The default is 'no date restriction', but it is also possible to restrict the search to only identifying documents that were published since a given year. This is useful in two situations. First, when you have an inclusion criterion that states that only papers published after a specific date are eligible. Second, when you want to update a search and only look for records that were published since the original search was conducted.
- It is possible to filter results using the Cochrane RCT Classifier. This is useful in situations where you are only interested in finding randomized trials. We have two options for the classifier: the 'high sensitivity' threshold, that is authorised by Cochrane, and a 'higher specificity' threshold, that will eliminate more non-RCTs than the other one.
- Finally, you can specify the 'search mode' that will be used. The seven available options are those outlined in Figure 2 and listed above in this section, i.e. you can retrieve 'new 'records that are connected via either or both 'citation' and 'related publications' ('composite network') connections, in either one or both directions.

When you have set up your search, you 'save' it and it will then appear in your list of saved searches. The server checks periodically for new searches (every two minutes) and runs new searches in sequence. There can therefore be a delay of up to a few minutes between submitting a search for

related records and the results appearing. You can refresh the screen by clicking 'auto-update home'.

Once the search has been run, its status will change from 'pending' to 'complete' (Figure 4). When a search has returned its results, you can either import the new records into your review or preview them – either within *MAG Browser* in EPPI-Reviewer, or via a link out of EPPI-Reviewer to the *Microsoft Academic* standard web UI (https://academic.microsoft.com/home).

Figure 4: Find related papers window with one 'complete' search



If you want to import the retrieved records, click 'import' and they will be brought into your review, and allocated to a new 'source'. The number of records imported may be smaller than the number on this screen because, if some of the records are already in your review, then these records will not be duplicated.

If you click the number of results (e.g. '296' in the example shown in Figure 4), you are taken into a screen that enables you to browse the *Microsoft Academic Graph*, with the items that your search found listed on the right. On the left are the topics (*MA* 'Fields of Study') associated with the records listed on the right.

While the purpose of this screen is mainly to enable you to preview your search results, it does contain a number of useful (standard) features (see below).

Using Standard Features in *MAG Browser*

1. Selecting records

It is possible to 'select' *MAG* 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 EPPI-Reviewer database, so it 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.

2. Browsing by topic

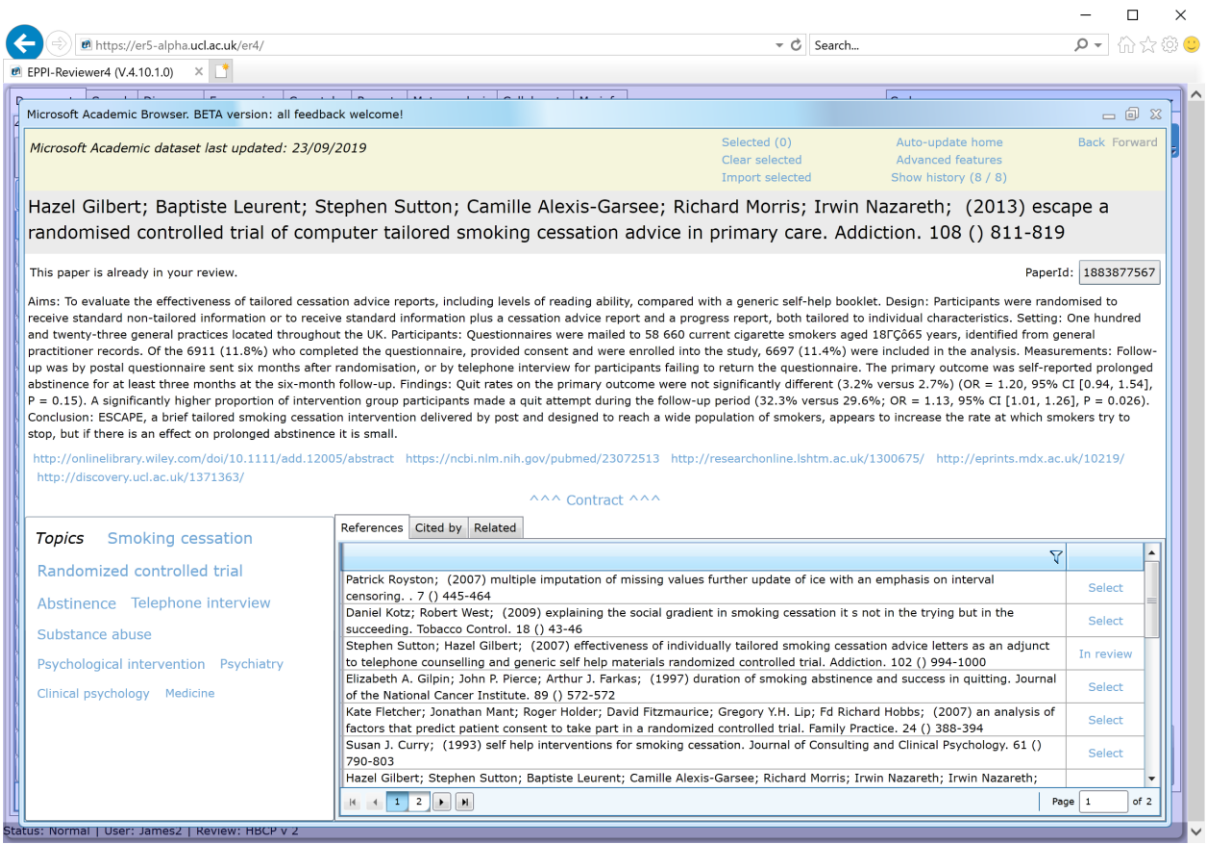
You can click the 'Fields of Study' topics on the left-hand side of the screen to examine the records that have been automatically classified as belonging to that topic by the *Microsoft Academic* algorithm. These topics are organised in a hierarchy (with 'broader' topic terms listed towards the

top), so you are able to navigate 'up' and 'down' this hierarchy. The 'Fields of Study' topics themselves have been generated automatically by *Microsoft Academic*, so they may change from time to time as and when the underlying algorithm is re-run on new data and the models are updated.

3. Detailed information about a specific MAG record

Figure 5 (above) shows a *MAG Browser* view of some of the detailed information about a specific document indexed in *MAG* (i.e. information in the *MAG* record). Each document (research article) may have been be found by *Microsoft Academic* 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 the selected document - including all identified full-text sources. This *MAG Browser* view also presents lists of the documents (*MAG* records) included in the selected document's bibliography, the documents that cite it, and those that are related to it according to *Microsoft Academic*'s 'Related Publications' ('Composite Graph') relationship algorithm (see above in this section). The 'Fields of Study' topics associated with the selected document, and whether or it is a 'study report' that is already included in your review (i.e. whether it is, or is not, among your current set of 'known' documents) are also shown.

Figure 5: Detailed information about a specific MAG record



Use Scenario 2: Continuous updating of a living systematic review

As described in relation to first use scenario (see above in this document), it is possible in *MAG Browser v1.0* to set up automated searches for 'new' (or 'unknown') documents that are connected

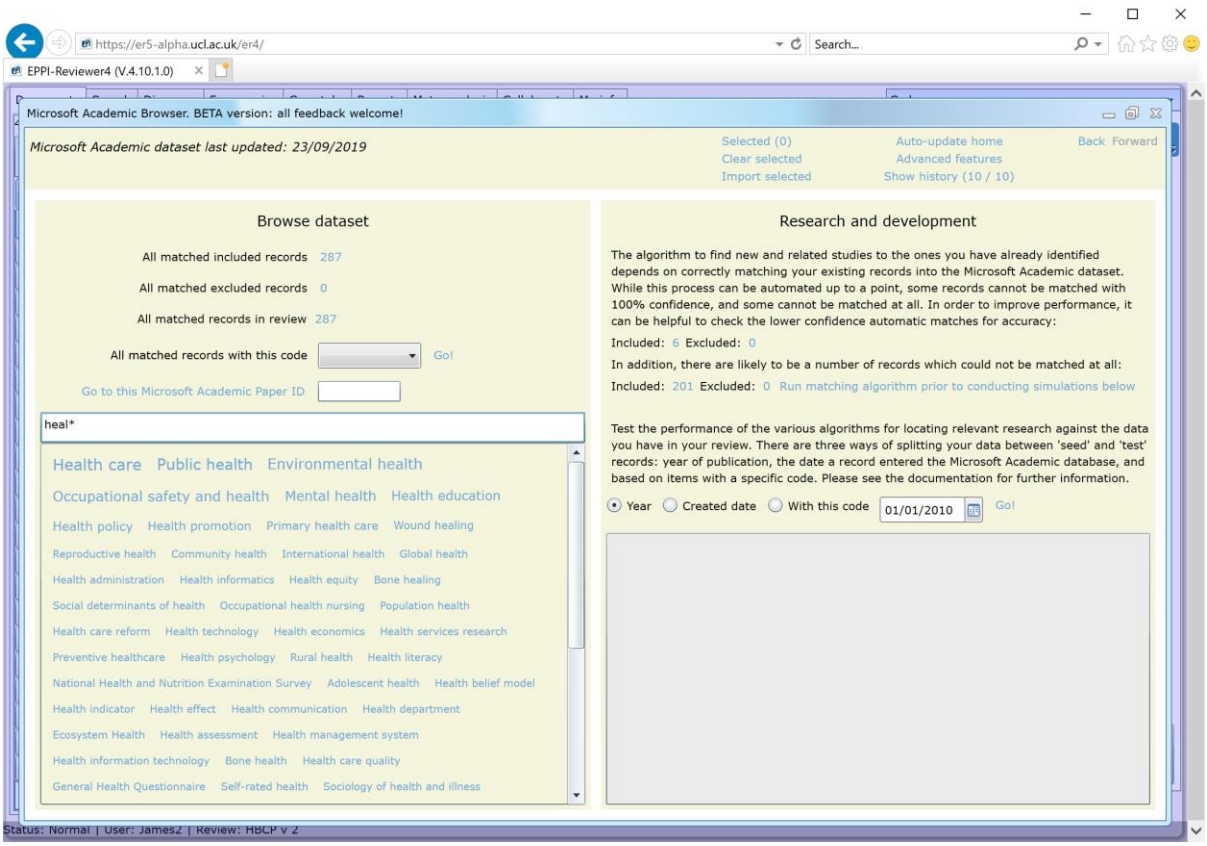
to 'known' documents already included in your EPPI-Reviewer review. This 'auto-update' search feature is designed primarily to support the continuous updating of reviews using 'living systematic review' workflows[5].

Every 10 days, an updated copy of *MAG* – which typically includes around half a million new records – is integrated into the 'back-end' of EPPI-Reviewer. If the 'auto-update search' checkbox is checked (Figures 3 and 4), then your search is re-run every time a new version of *MAG* becomes available. New records that are connected to those already in your review (using the selected 'search mode'/ network graph relationship(s) from seven current options – see above) will be identified; and a new row containing those new records is added to the list of searches shown in Figure 4.

Using Advanced Features in *MAG Browser*

There are two distinct types of advanced features in *MAG Browser v1.0*: a) Browse the *Microsoft Academic Graph*; and b) Research and Development. Both of these can be accessed by clicking the 'Advanced features' link at the top of the screen.

Figure 6: Advanced features



[5] For further information about 'living systematic review' methods, tools and workflows, please visit: https://community.cochrane.org/review-production/production-resources/living-systematic-reviews.

Browse the *MAG*

It is possible to 'Browse the *Microsoft Academic Graph*' starting with those records already included in your EPPI-Reviewer review. If you have already identified a *MAG* record ID number that you are interested in, you can enter it into the box here and navigate to the record page. You can also search for topics using a free text search that can include the * wild card character (Figure 6).

Research and development

As outlined above in this document, we are releasing *MAG Browser v1.0* as a new suite of tools that requires further research and development in order to optimise their use for efficient study identification in systematic review, evidence synthesis and other use scenarios. We therefore strongly encourage users who are interested in using the functionality described above to rigorously evaluate the use of *MAG Browser* to support study identification in their own review(s) and other use cases. Please share your evaluation findings and feedback with the EPPI-Centre team, via e-mail to: EPPISupport@ucl.ac.uk – including 'MAG' in the e-mail subject line.

The 'Research and Development' feature in *MAG Browser* is designed to support such evaluations in various use scenarios; and the information below describes how this advanced feature - and its integrated *MA Graph Search* feature (see below for details) – can be used for this purpose. We have also produced a flyer on 'Using *MAG Browser and MA Graph Search* for Research & Development in EPPI-Reviewer', available from the *MAG Browser Portal* at: https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3754.

First, use the 'Research and Development' feature to curate a 'gold standard' unified dataset. This involves automatically matching 'known' records (e.g. study reports included in the current version of your review) to corresponding *MAG* records and then:

1. View + check 'auto-matched' records;
2. View + check 'close match' records; and
3. View, manually look-up + code 'unmatched' records

Checking 'auto-matched', 'close match' and 'unmatched' records

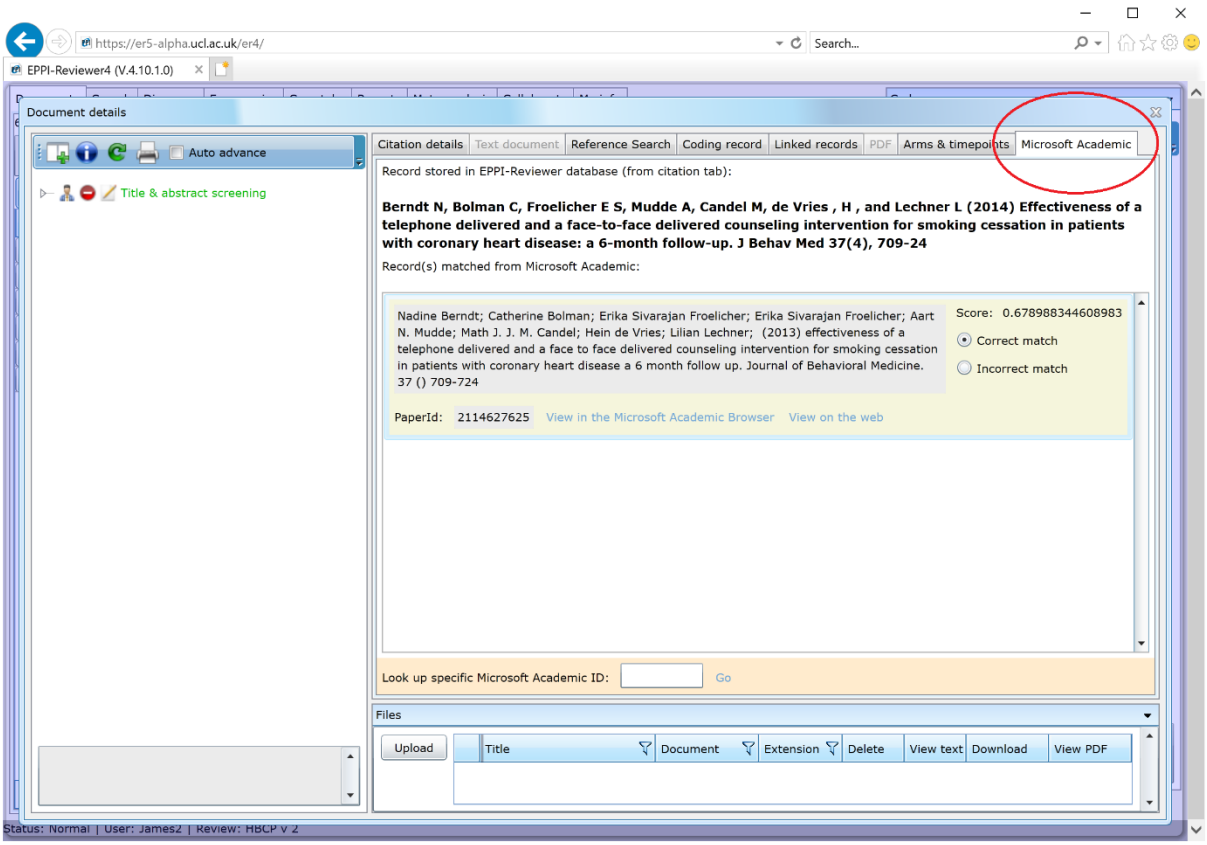
The success of the automated matching procedure – and the resulting 'completeness' and 'accuracy' of the unified dataset – depends on the ability of our 'matching algorithm' to accurately match the selected set of 'known' documents already included in your EPPI-Reviewer review with their corresponding *MAG* records. We have developed this 'matching algorithm' carefully and it is highly accurate, but there will be cases where the match is either uncertain (i.e. 'close matches') or can't be made at all (i.e. 'unmatched'). In this scenario, the accuracy of the matching will be improved if the 'close match' and/or 'unmatched' records are manually checked by the user. Ensuring that 'matches' and 'non-matches' are correctly labelled is also important for robust evaluations of *MAG*, the *MAG Browser* and its features.

'Close match' and 'unmatched' records can each be listed on the home page of EPPI-Reviewer by clicking on the 'numbers' of each type of record showing in the 'Research and Development' pane. Once listed on the EPPI-Reviewer home page, you can go into the 'coding details' window of each record (by clicking 'Go'); go to the *Microsoft Academic* tab on the right-hand side; and see the

'known' record in EPPI-Reviewer (in bold at the top of the screen) and the MAG record(s) which have been automatically matched below it. You can then assess whether or not the match is correct and save your judgment by clicking 'correct' or 'incorrect' match.

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 search for a corresponding *MAG* record manually via the Microsoft Academic standard web UI (https://academic.microsoft.com/home) – which can be opened from the *MAG Browser* – and then, if a corresponding *MAG* record is found, you can enter the appropriate *MAG* record ID number in the *Microsoft Academic* tab on the 'coding details' window, add the match manually and finally mark it as a 'correct' match.

Figure 7: Checking and adding matches between review records and *MAG*



Simulating updates

In the 'Research and Development' pane, it is also possible to simulate the performance of the seven currently available options for generating *MA Network Graphs* to identify 'new' ('unknown') records by splitting records from a completed review into two sets. For example, consider a review that was completed in 2019, containing records from the previous couple of decades. We could simulate the scenario where the review was actually completed in 2015, and subsequently updated in 2019 by splitting the records into two sets: those published before 31 December 2016 and those published since. We can then look to see how many of the newer records can be found using *MA Network Graph* relationship(s) connections from the older ones. Publication year might be insufficiently granular to facilitate good evaluation, so it is also possible to use the date on which a record first appeared in *MAG* to 'split' the records into two sets. Please note that as the dataset is relatively

new, most of the records have a created date of 24th June 2016, so it is only possible to use this date to simulate quite recent updates using this option.

Simulation results are shown in the lower portion of the 'Research and Development' pane. Data outputs include the simulated 'recall' and 'precision' of each of the seven current options for generating and using *MA Network Graphs* to identify 'new' ('unknown') documents.

MA Graph Search algorithm

We have also developed a novel *MA Graph Search v1.0* algorithm, designed to increase the precision of *MAG* searches, using any of the seven current options for generating and using *MA Network Graphs*. This algorithm currently utilizes a concept drawn from graph analysis – one of 'degree centrality' – which measures how 'connected' a given 'node' in a graph is. Our current theory is that documents that are more 'connected' to others (via semantic, citation, authorial and institutional relationships) in our graph search are more likely to be of interest. Early evaluations of this approach are promising, but more research is needed. (Facilitating research & development in this area is a primary reason for us making this feature available at this time.)