Development of a prototype tool to automatically translate literature search syntax

Jennifer Hill (a) and Cong Chen (b)
(a) All Hazards Public Health Response Evidence Review Team, UK Health Security Agency, London, UK

Abstract
This article describes the development of a prototype tool to automate the translation of bibliographic database search strategies from the Ovid platform to a range of other platforms. The development of this Bibliographic Syntax Converter (BSC) took place as a collaboration between an information specialist and a team of coders during a two-day Hackathon. This collaboration illustrates the potential benefits of this kind of cross-disciplinary working. A discussion of the difficulties inherent in automating the translation of literature search syntax is also provided, using specific examples to demonstrate these difficulties. The article concludes with a brief consideration of the limitations of the BSC tool, and plans for future development.

Key words: *information storage and retrieval; library science; library automation; systematic reviews.*

Background

The UK Health Security Agency (UKHSA) Knowledge and Library Services (KLS) provides evidence support services to UKHSA staff and Local Authority Public Health teams in England, including literature search services. Literature search requests require detailed systematic searching of bibliographic databases, usually across several different databases. UKHSA KLS receives a high volume of search requests, 400 in 2021-22. UKHSA's remit is to protect the health of the public against infectious diseases and other public health hazards. The evidence base for public health is widely distributed across different domains (1) and searching of databases covering a range of topics is necessary (2). Literature searching requires translation of search strategies to databases hosted on different platforms, which use different syntax. This variation across platforms means Information Specialists carrying out searches spend time manually re-entering search terms into each new database and platform to be searched, editing operators for each platform as they go.

Automation and machine learning tools are becoming increasingly popular for aiding the conduct of systematic reviews, as they offer time saving efficiencies. One estimate suggests that there are around 160 existing tools intended to help with one or more stages of the review process (3), and another lists 235 tools (4). Currently there are few existing tools which automate syntax translation such as replacing the proximity operator for the Ovid platform with the correct equivalent for other platforms such as Web of Science or EBSCO. Two tools that perform this function are PolyGlot Search, as part of the Systematic Review Accelerator tool (5) and Medline Transpose (6). Another method involves using macros in Word documents for translation (7). PolyGlot Search provides capability to translate searches between databases including Ovid MEDLINE, PubMed, Web of Science, Scopus and others. Medline Transpose translates strategies between Ovid MEDLINE and PubMed only. Whilst these tools can be used for automated translation of search syntax, there are limitations. PolyGlot Search is, at the time of writing, unable to carry out any translation of subject headings/thesaurus terms between databases. This is an undeniably challenging task given the differences in indexing terms available in different databases (especially where databases relate to different subject areas), the fact that thesaurus terms are regularly updated (annually for the MeSH thesaurus)

Address for correspondence: Jennifer Hill, All Hazards Public Health Response Evidence Review Team, UK Health Security Agency, 10 South Colonnade, Canary Wharf, London, UK. E-mail: Jennifer.hill@ukhsa.gov.uk

and need for human judgment in determining the best subject heading to use in cases where an exact or very close equivalent index term is not available.

In November 2022, UKHSA held a two-day Hackathon across the organisation. This provided an opportunity for KLS staff to work with data scientists, exploring the possibility of developing a Bibliographic Syntax Converter (BSC) tool to automate conversion of search syntax. Participation in the Hackathon provided a learning opportunity for KLS staff to understand more about applications of data science to automate search tasks and to gain experience of working with coders to develop solutions. Whilst development of a tool capable of translating thesaurus terms between databases was beyond the scope of the two-day event, basic syntax translation was a necessary first step to enable thesaurus translation in the future.

The aim of this paper is to describe the development of a new tool to automate the translation of literature search syntax from the Ovid platform to other platforms, including EBSCO, Web of Science, ProQuest etc., and to illustrate the benefits of cross-disciplinary working between data scientists and information specialists.

Methods

The BSC tool was designed and developed in November 2022 and was created to accept a search strategy written in Ovid syntax. The tool was developed through collaboration between UKHSA data scientists and an information specialist.

The tool was designed to convert syntax from the Ovid platform to EBSCO, ProQuest, Cochrane Library, Web of Science and Scopus. It is important to note that the tool translates syntax between platforms, not databases. Therefore, the tool translates a search from Ovid syntax to EBSCO syntax rather than from Medline syntax to CINAHL syntax. Syntax is determined by platform rather than database, i.e. Medline searched via EBSCO requires use of EBSCO syntax while Medline search via Ovid requires use of Ovid syntax. A platform uses the same syntax throughout for search operators, wildcards and truncation symbols, meaning that these elements of a search would be the same for any EBSCO search, whether the search was being conducted on CINAHL, Global Health or any other database via EBSCO platform. For this reason, the decision was taken to translate searches between platforms, focusing on translating elements of syntax that are common across platforms. Those platforms most frequently used by UKHSA KLS were chosen for inclusion in the tool. *Figure 1* provides an overview of the input and output of the tool.

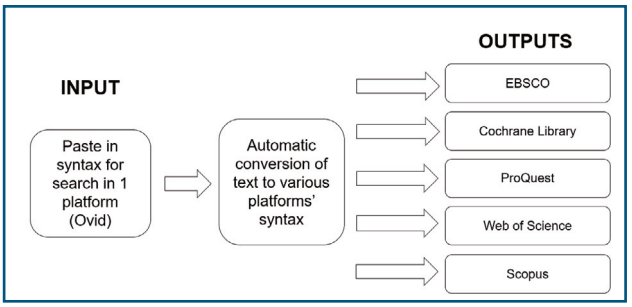

Fig. 1. *Overview of bibliography syntax converter.*

Elements of syntax which the prototype BSC tool could translate were:

- common two letter Ovid field codes e.g. .tw, ti, ab, kw, kf.
- proximity operators
- optional wildcards (wildcards that can stand for 1 or 0 characters in a word) and mandatory wildcards (wildcards that must replace 1 character in a word)
- truncation symbols
- Ovid command line syntax for combining multiple lines of search (i.e. or/1-5)
- AND and OR operators
- requirements for quotation marks in phrase searching for relevant database platforms.

Technical aspects of development

Python was chosen due to experience across the organisation ensuring support for future development and deployment. To facilitate processing and human review of complex strategies it was important that the tool should accept an exported text copy of the search strategy and provide a line-by-line translation, with line-by-line warnings where the tool was unable to translate elements of the search strategy.

Specifics of mapping (e.g., which combinations of fields should be mapped) were held in text format so users could easily make improvements. The mapping process for each search strategy line broke it down as the writer might, with low level thinking for strings and wildcards in simple searchers and higher levels of interpretation for combinations with operators or com-

binations of fields. These aimed to take the search strategy into a form people could reason about and verify. This intermediate form could then be translated back into different syntax representations, made easier by separating concerns between the different levels.

Collaboration between information specialist and data scientists

Having outlined the overall design, the information specialist was closely engaged throughout the development process to ensure that mapping details were clearly exposed and user-updatable and that functionality was prioritised to meet user needs. To provide a detailed description of the process of developing the BSC tool, examples of the factors that were considered during development are provided below. This illustrates the input provided by the information specialist and some of the decisions that had to be taken to allow data scientists to create a program that would work to the specifications of information specialists at UKHSA.

Example 1: Correctly matching proximity operators

Most of the commonly used platforms such as Ovid, EBSCO, ProQuest and Web of Science allow the user to search several different databases. Some elements of search strategy syntax are consistent across the platform and can therefore be used in any database that is searchable through that platform. For example, the same proximity operator can be used to search any database via EBSCO platform. However, proximity operators between platforms differ. These differences affect not only the actual text used for proximity operators in each platform, they can also affect the rules by which operators are applied in a platform, and how unqualified proximity searches (searches where a number is not specified with the proximity operator) are handled.

To enable data scientists to write a program that would accurately translate proximity searches required the information specialist created a table detailing how operators should be mapped (*Table 1*).

Ovid (Medline/Embase)	EBSCO	Cochrane	Web of Science	Scopus	Proquest
Risk adj3 assessment	Risk N2 assessment	Risk NEAR/2 assessment	Risk NEAR/2 assessment	Risk W/2 assessment	Risk NEAR/2 assessment
Risk adj assessment	Risk N1 assessment	Risk NEAR/1 assessment	Risk NEAR/1 assessment	Risk W/1 assessment	Risk NEAR/1 assessment
Risk adj16 assessment	Risk N15 assessment	Risk NEAR/15 assessment	Risk NEAR assessment	Risk W/15 assessment	Risk NEAR/15 assessment
Risk adj7 assessment	Risk N6 assessment	Risk NEAR/6 assessment	Risk NEAR/6 assessment	Risk W/6 assessment	Risk NEAR/6 assessment
Risk adj5 assessment	Risk N4 assessment	Risk NEAR/4 assessment	Risk NEAR/4 assessment	Risk N/4 assessment	Risk NEAR assessment

Table 1. *Mapping of proximity operators between database platforms.*

Example 2: Deciding the most efficient way in which to map fields from Ovid MEDLINE to other databases

Searchable fields can differ widely between the databases available through a platform, presenting a challenge in terms of coding the syntax converter. Due to time constraints of the Hackathon, the tool was developed to translate searches between platforms rather than specific databases. This meant that a full translation of all fields available in a specific database such as Ovid Medline was not attempted. Instead, the information specialist working on the project proposed focusing on translating fields that are common to all platforms required (title, abstract and keywords). These fields would be considered essential for advanced systematic searches suitable for systematic reviews (8). Preferred options for mapping these fields in Ovid were provided to data scientists, along with details of the differences in formatting of search fields across platforms (*Table 2*).

Example 3: Truncation, wildcard and phrase searching

A third element to consider was the translation of truncation, wildcards and use of quotation marks (*Table 3*). One of the most commonly used syntax elements within this area is the right-hand truncation applied at the end of the root part of a word in order to search multiple variant endings. The asterisk symbol can be used for this purpose across all platforms which the BSC deals with.

Mapping is also relatively simple for mandatory and optional wildcards.

While symbols used differ in some platforms, the BSC tool only needs to swap the Ovid symbol for the appropriate symbol in each other platform.

Fields searched	Ovid	EBSCO	Cochrane	Web of Science (Advanced search)	Scopus	Proquest
Unqualified/default field searches	No field code required	No field code required	No field code required	No direct equivalent – fields must be specified	No direct equivalent – fields must be specified	No direct equivalent – fields must be specified
Title and abstract	.tw .ti,ab	No direct equivalent without repeating search terms – leave unqualified	No direct equivalent without repeating search terms - leave unqualified	TS=(risk assessment)	TITLE-ABS(risk assessment)	ABSTRACT,TITLE (risk assessment)
Title	.ti	TI xxxxx	:ti	TI=(risk assessment)	TITLE(risk assessment)	TITLE(risk assessment)
Title, abstract and keywords	.tw,kf	No direct equivalent without repeating search terms – leave unqualified	No direct equivalent without repeating search terms – leave unqualified	TS=(risk assessment)	TITLE-ABS-KEY(risk assessment)	ABSTRACT,TITLE ,IF(risk assessment)

Table 2. *Mapping of search fields and formatting across platforms.*

Symbol	Ovid	EBSCO	Cochrane	Web of Science	Scopus	Proquest
Optional wildcard – 0 or 1 characters within a word	Tumo?r	Tumo#r	Tumo?r	Tumo$r	Not available	Tumo*r
Mandatory wildcard – 1 character within a word	Organi#ation	Organi?ation	Not available	Organi?ation	Not available	Organi?ation
Right hand truncation, 0 or more characters	Risk assessment*	Risk assessment*	Risk assessment*	Risk assessment*	Risk assessment*	Risk assessment*
Left hand truncation, 0 or more characters	Not available	Not available	*flight	*flight	Not available	Not available

Table 3. *Mapping of wildcard and truncation symbols.*

Limitations of the BSC and future development

The syntax converter outlined in this article was created within a limited period as part of a two day Hackathon event in late 2022. The time constraints meant that the first iteration of the tool was limited to converting searches designed for Ovid MEDLINE to outputs suitable for use in a limited number of platforms. At the time of writing, the tool does not allow translation of syntax from any other platform apart from Ovid.

Another limitation is that the tool is not currently able to translate Ovid searches using multiple fields. For example, an Ovid search term such as "risk assessment".tw,kw would search title, abstract and keyword fields. However, the prototype version of the BSC tool will only read and translate the first two letter field code from this search term. The ',kw' portion of the search is not translated, and this failure is flagged in the outputted search strategies. For databases such as Web of Science and Scopus, a searcher may want to search the TS or Title-Abstract-Keyword fields. The .tw. from Ovid can be mapped directly to these more inclusive alternatives, however this would result in a loss of precision to the search due to the additional inclusion of keyword fields in TS. In addition, the tool is currently only able to translate searches between platforms rather than between databases. In order to introduce translation between specific databases such as from Ovid Medline to EBSCO CINAHL, further work would be needed to expand the number of search fields which can be translated.

The tool is currently also only able to translate text parts of search strategies by swapping field codes, truncation symbols and search operators to the closest equivalent in each platform. It would enhance the benefits of using a tool to automate translation if it could also be developed to convert subject heading terms, or at least to provide a searcher with a list of potentially relevant equivalent subject headings to choose from. This aspect of development is more complicated and may require use of database platform APIs. Plans are in place to explore whether this functionality could be added to the tool, but this will require further coding resource and support.

Minor developments planned include introduction of mapping for a greater range of fields beyond the title, abstract and keyword fields included initially. Given the range of fields available in some databases, it will be necessary to prioritise selection of fields to map first. This and other developments will be identified through testing and evaluation of the BSC tool. UKHSA KLS has set up a working group to take forward development of the tool, and one of the first tasks will be for group members to begin using the tool in daily workflows in order to identify problems with existing functions and additional functions that would be helpful to add.

Conclusions

The development of the UKHSA's prototype BSC converter tool within a limited two day time-frame shows how much can be achieved relatively quickly in terms of automating library and information science workflows. Given the volume of literature searches which KLS carries out each year the opportunity for time-saving benefit is clear. These benefits can be further enhanced through developments to the tool, along the lines suggested above. As a first next step, an internal working group has been set up in order to carry out thorough testing of the tool. The intention is that this testing will highlight additional, as yet unidentified areas for development and may also pave the way for a more formal evaluation of the benefits of using the tool to automate this aspect of information retrieval.

Received on 29 July 2024.
Accepted on 20 November 2024.

REFERENCES

1. Heath A, Levay P, Tuvey D. Literature searching methods or guidance and their application to public health topics: A narrative review. Health Information & Libraries Journal. 2022;39(1):6-21.
2. Hanneke R, Young SK. Information sources for obesity prevention policy research: a review of systematic reviews. Systematic Reviews. 2017;6(1):156.

3. Cierco Jimenez R, Lee T, Rosillo N, Cordova R, Cree IA, Gonzalez A, Indave Ruiz BI. Machine learning computational tools to assist the performance of systematic reviews: A mapping review. BMC Medical Research Methodology. 2022;22(1):322.
4. Johnson EE, O'Keefe H, Sutton A, Marshall C. The Systematic Review Toolbox: keeping up to date with tools to support evidence synthesis. Systematic Reviews. 2022;11(1):258. doi: 10.1186/s13643-022-02122-z.
5. Clark JM, Sanders S, Carter M, Honeyman D, Cleo G, Auld Y, et al. Improving the translation of search strategies using the Polyglot Search Translator: a randomized controlled trial. J Med Libr Assoc. 2020;108(2):195-207.
6. Wanner A, Baumann N. Design and implementation of a tool for conversion of search strategies between PubMed and Ovid MEDLINE. Research Synthesis Methods. 2019;10(2):154-60.
7. Bramer WM, de Jonge GB, Rethlefsen ML, Mast F, Kleijnen J. A systematic approach to searching: an efficient and complete method to develop literature searches. J Med Libr Assoc. 2018;106(4):531-41. doi: 10.5195/jmla.2018.283.
8. Cochrane. Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022)2022.