Matching Data from Heterogeneous Databases for Integrated Assessment of Research Productivity

C.J. Tran

Computer Science, Class of 2023

Honors Student

Advisor: E. Andrew Balas MD, PhD. Professor

Biomedical Research Innovation Laboratory, Augusta University

In-Field Reader: Clément Aubert, Assistant Professor of Computer Science

Honors Advisor: Josefa Guerrero Millan, Associate Professor of Physics

Correspondence concerning this article should be addressed to C.J. H. Tran, Augusta University CJ2308 (IPPH) 1120 15th Street Augusta, GA 30912, United States. Email: 2cj.tran@gmail.com Phone Number: 803-271-5472

INTRODUCTION

When research is being conducted, new information continually populates multiple and diverse research databases. Eventually, information relevant to the same unit (e.g., person or organization) will exist across many national and international databases. Meanwhile, they use various and often very different identifiers, so analyzing and combining data from different databases can be difficult.

To address this issue, the field of research in matching data across databases and merging multiple datasets has become a blossoming field of study, with some preliminary research already existing in more specialized fields or with a focus on the techniques to match and merge data—in particular, utilizing machine learning (e.g., a fuzzy search algorithm).

For example, in life sciences research, a large variety of results and research are produced: scholarly articles, books, patents, genetic clinical experimentations and protein sequences, practice recommendations and others. Furthermore, there are millions of academic articles, with more being published yearly (Landhuis, 2016). These articles are indexed in various types of databases. Naturally, there would be data about a research laboratory in multiple publication and grant databases that have no universal identifier to match these data from different sources.

This issue in the field of research has been documented before, like in Halpin and McNeill's research investigating heterogenous data sources, where they noted how data sources are vast in size, continually expanding, ripe for academic collaboration, but due to variations in "terminology, structure, and formats" used, it becomes difficult to even "ascribe meaning to the terms of that data source, [much less]… understand what is being conveyed" (Halpin & McNeill, 2013). Similarly, Christen established data matching as a real issue, providing background context and relevant terminology associated with data matching (e.g., "record linkage, entity resolution, object identification, duplicate detection, identity uncertainty, [and] merge-purge"), as well as noting the main challenge with data matching to be how identifiers may not be consistent or "available in the matched databases" (Christen,

2014). Meanwhile, Pang et. al investigates data matching in the context of preserving the privacy of user information while also utilizing a matching methodology like fuzzy matching, a method to match similar data (Pang et. al, 2009). Furthermore, de Leeuw and Keijl investigated the discrepancy between multiple database use and the noted connections between databases, which leads to difficulties in replicating studies, so they provide informative decisions to help other researchers match databases (de Leeuw, & Keijl, 2015). Not only that, Lara et. al noticed in their study over accounting and financial firms that empirical results vary depending on sample choice, despite using the same research design, allowing them to conclude how database choice matters (Lara et. al, 2006).

Unlike those studies, our study seeks to create a more general framework for developing a matching algorithm, with an emphasis on identifiers and vocabularies when matching. The purpose of this study is to develop a method and an algorithm, which can match information about the same unit (e.g., university, researcher, or journal, etc.) from a large variety of differently structured databases so that better analysis can be performed over research productivity.

COMPUTATIONAL METHODOLOGY

To understand research productivity, it is necessary to create a matching algorithm that retrieves divergent database information to get a comprehensive data about selected units of analysis and create a primary dataset for further study.

Term	Definition
Unit of analysis	The entity (e.g., person or organization) searched in a database for information
Data source	A specific collection of data from which information can be gathered
Synonym/Identifier of the unit of analysis	A variable that specifies the unit of analysis
Search strategy	A unique combination of identifiers and techniques, used to find information in a database
Raw dataset	A dataset of information extracted and output from a database using specific search strategies
Integrated dataset	The curated compilation of selected raw data for analysis
PubMed	"A free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health-both globally and personally…" (NCBI, 1996)
Google Scholar	"One place… you can search across many disciplines and sources… [to help] you find relevant work across the world of scholarly research" (Google, n.d.)
NIH RePORTER	"… an electronic tool that allows users to search a repository of both intramural and extramural NIH-funded research projects and access publications and patents resulting from NIH funding." (NIH, n.d.)
USPTO (The United States Patent and Trademark Office)	"… the federal agency for granting U.S. patents and registering trademarks" (USPTO, 2018)

Figure 1. Definitions of terms

STEP 1: SELECT A UNIT OF ANALYSIS

Description

To begin the matching process, choose a specific unit and establish knowledge of what will be searched for. This unit is simply a flexible term to represent any singular, basic item that will be searched for in a database. This could range from a researcher name to a university, or the title of research—so long as the item can be searched for in a database (**for example**: Researcher – John Doe). Knowledge of what to search the databases for is also required, since it would be inefficient to aimlessly search a database and ineffective if the data found in a database is not related in some way**.**

STEP 2: IDENTIFY INITIAL IDENTIFIERS

Description

To begin the step to start and search multiple databases, an initial set of identifiers will need to be manually chosen based on the needs of the particular project. Ideally, matching data between databases is as automated as possible, however, to begin the matching process, there is some degree of manual input required to at least start the matching process. This initial set of identifiers that will be manually input will serve to recognize the unit of analysis and help find other identifiers in other databases. **For example,** an initial set of identifiers include name, email, degrees, positions, employing institute, city, state, country, NIH contact PI ID, etc.

STEP 3A: CREATE SYNONYM VOCABULARIES

Description After the initial set of identifiers are created, while searching databases and matching data, similar identifiers to the initial set of identifiers will be encountered. These identifiers can be grouped together with the initial set of identifiers to form synonym vocabularies, which are essentially groupings of related identifiers. These synonym vocabularies allow for more improved matching between different identifiers, since new synonyms should be compared to the entire vocabulary or group (since they are all related). **For example,** "John Doe", "Doe, John William", "John W. Doe", "Doe JW", and "Doe, John W." are all in a synonym vocabulary because they are similar enough to be related yet simultaneously not exactly the same—whether this is how they are spelled, arranged, abbreviated, etc.

These synonym vocabularies can be created (i) automatically, (ii) manually, or (iii) both, can be (i) unambiguous; (ii) confirmed/additionally verified/likely; or (iii) unconfirmed. An aid to help create a synonym vocabulary could be to incorporate an online database or website to identify root words or common abbreviations. Two aids seemed helpful: (i) LTWA (The List of Title Word Abbreviations): a

standard for abbreviations; (ii) ISO 4: a substandard of LTWA specifically used to abbreviate the names of scientific journals. However, they are still conventions and may not consider every scenario; they're only useful if the abbreviation follows convention/rules, so misspellings or abbreviation variations may not be accounted by such strict conventions.

STEP 3B: IDENTIFY SECONDARY IDENTIFIER CANDIDATES

Description After the initial set of identifiers and their associated synonym vocabularies are developed, secondary identifier candidates can be detected. Secondary identifier candidates, similar to the initial set of identifiers, are identifiers that are relevant to the unit of analysis and are used to help further find or verify specific information in the database. **For example:** ID, city, age, publications, department, or organization type.

STEP 4: DATABASE SEARCHES

Description Now that the identifiers and vocabularies have been established, the databases need to be searched for information. **For example**, these databases could be PubMed, Google Scholar, NIH RePORTER, or USPTO.

STEP 4A: GENERATE SEARCH STRATEGIES

Description To do this, the identifiers and vocabularies created can be combined in unique combinations to comprise a search strategy, or how information will be searched for in a database. While there can be a myriad of combinations that can be made, typically, they will be specifically tailored to a database since identifiers and vocabularies often differ between databases. **For example,** {Name: "John Doe", Emory University} could be used to search the NIH RePorter database.

STEP 4B: SEARCH AND EXTRACT

Description: Naturally, after the search strategy is developed, the next step is to utilize the search strategies to search the chosen data sources for information. Once information is found, the search result should be extracted into a raw dataset—ideally, in the form of an Excel structure for simpler data matching and analysis. Afterwards, the relevant information is extracted into an Excel file—each data source typically has a button that allows for the extraction of data—with those files serving as the raw dataset. **For example,** the subsequent data sources are searched using the subsequent search strategies/identifiers: (i) PubMed: John W Doe, Doe JW; (ii) Google Scholar: John Doe, JW Doe; (iii) NIH RePORTER: Doe, John William, Doe, John W. ID: 12345678; and (iv) USPTO: Doe, John W.

STEP 5: MATCH DATA

Description: Now that the data is extracted from the data sources into raw datasets, the data needs to be combined into one integrated dataset to be used for analysis and further research. This collation of data can be done through matching. However, when matching data, there exists different matching scenarios we classified as: identical matches, root words, using other identifiers, and manual matching.

OPTION 5A: IDENTICAL MATCHES

Description True to its name, identical matches occur when identifiers or vocabularies of data exactly match between datasets. Identical matches are not difficult to match, but unfortunately, are rarely occur in the real world (**for example**: “John W Doe” and “John W Doe” are exact matches).

OPTION 5B: ROOT WORDS

Description On the other hand, data may not identically match, but may share a common root. Roots are the structural building blocks from which words obtain their meaning. Better defined, they are “the

simple element inferred as the basis from which a word is derived by phonetic change or by extension (such as composition or the addition of an affix or inflectional ending)" (Merriam-Webster, n.d.). However, abbreviations also exist, which share a similarity with roots in how they both retain the basic element of a word while still being able to gain extensions. Therefore, for the purposes of our study, we combined the definition of roots and abbreviations into the term "root words" (**for example**: "biotechnology" and "biotech", "institution" and "institut-", "medicine" and "med", "agricultural" and "agric-", "university" and "univ").

To aid in the matching process of root words, there are some noteworthy observations that exist when comparing words. For one, if the entire phrase of one identifier is within another identifier, then the phrase likely matches (**for example**: "University of Research" in "University of Researches", "University of Research.", "University of Research in Study of Journals", etc.). Secondly, if some words of a phrase match—but not all—then check the unmatched words for root words and if there is a shared root word, then the entire phrase likely matches (**for example**, with "School of Science" and "School of Sci", the phrase "School of" match but then you can also see "Science" and "Sci" have the same root word). Alternatively, if an exact abbreviation is used, the full term could be substituted, and the matching could be recomputed (**for example**: "USA" and "United States of America" when comparing "Library of USA" and "Library of United States of America").

OPTION 5C: OTHER IDENTIFIERS

Description The next matching scenario is using other identifiers and, like identical matches, are true to their name since they utilize other identifiers to aid with matching data. After all, matching isn't limited to phrases and those with abbreviations; you can use other secondary identifiers as well to verify that certain data match. **For example**, your primary identifier could be the first name "John Smith," and while this identifier narrows the search in each dataset, there can be multiple people with this name, so

other identifiers can be used, like institution “University of ABC,” which can vastly narrow down the search, since it is less likely for a large number of people to have the same name to be in the same location (see Figure 2 for more examples). This can be done on various identifiers and with multiple identifiers to reduce error. That said, each identifier varies in usefulness, so some combinations of identifiers may be more effective than other identifiers for matching. **For example**: on their own, DOIs are a unique identifier and were made to always give exact matches, while titles and researcher names are very accurate and should usually give an exact match (since the title of a study or the name of an individual will usually be enough to determine an individual, though it is still possible for there to be duplicates), while year and publication are inaccurate identifiers that will not help specify certain data (since publications can output thousands of studies a year and millions of studies can be output in just one year).

Name	Example
Digital Object Identifier/DOI	doi: 10.1016/j.jhep.2020.09.031.
Uniform Resource Locator/URL	https://pubmed.ncbi.nlm.nih.gov/33038433/#affiliation-5
Title	“COVID-19: Discovery, diagnostics and drug development”
Core NIH Project Number	U01AI027196
Patent Number	4681933
Researcher/Author	John Doe
Organization	Emory University
Funding Amount	$20,000
Volume	74
Issues	1
Pages	168-184
Publication/Journal	Journal of Hepatology
Primary Agency	NIH
Year	2020

Figure 2. Illustrative Identifiers

OPTION 5D: MANUAL MATCHING

Description Finally, there is manual matching, which is where an individual will individually review and match data in datasets. Unlike matching algorithms, manual matching done by people will be far more accurate due to having the cognitive capacity to be flexible or make inferences, as opposed to the restrictive rules matching algorithms follow. This is particularly useful for situations in which matching data has no obvious correlation (**for example**, if an institution changed its name multiple times: "Research University", "Medical College of Location", "Study University", "School of MCL"). That said, manual matching should be used as a review of dataset information rather than the default or initial step, since it would be infeasible to sort the vast amount of data by hand due to humans being more time consuming and less efficient. Some tips to help with matching include sorting the data alphabetically (so the comparisons could be done with similarly spelled data near each other) or ignoring human input variation (**for example**, capitalization, symbols, etc.) since they could be a result of human error or different conventions (though the actual values of the comparisons should not be changed).

STEP 6: CURATE INFORMATION

Description Finally, after being extracted from the data sources, the collected and matched data from the raw datasets are combined into a single, integrated research dataset that could be analyzed by researchers of various fields (**for example**, the final integrated dataset may look like: Master ID: 34, 56, 77; Patent: A5, L6, G8; Grant Awards: $100, $20, $78, or see Figure 3 for a visual table representation). While the merged dataset may be mismatched due to different databases having differing amounts of information, it would have collated all the information from the various databases, making analyzation and future research easier.

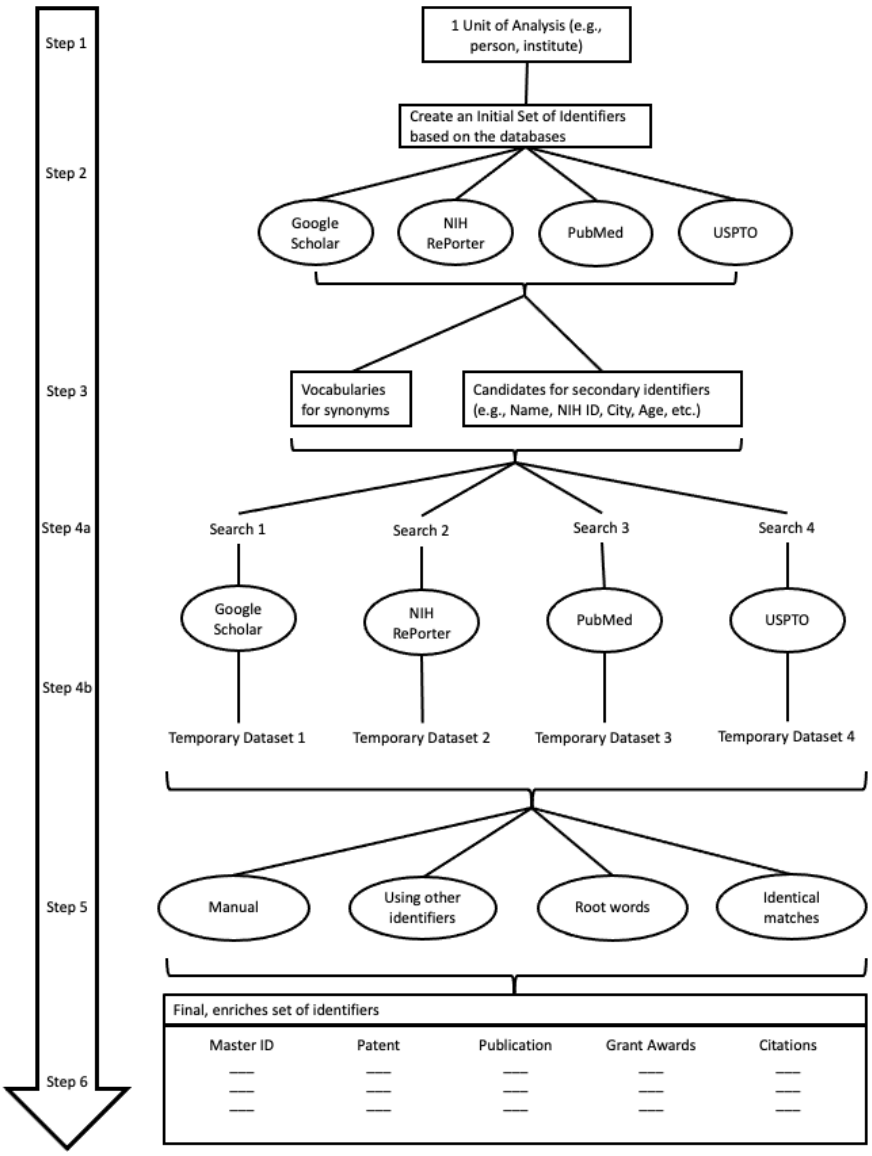


Figure 3. A visual representation of the matching methodology

RESULTS

To provide an accurate example and depiction of the search strategies in use, a real researcher is used. The following researcher (Stephen P. Bell, Massachusetts Institute of Technology/MIT) was manually searched for in each of the following databases (PubMed, Google Scholar, NIH RePORTER, and USPTO). Note the subtle differences in the search strategies for each database.

To start, after navigating to the PubMed website, “Stephen Bell” was entered into the search bar. This returned a few some search results and examining some of the top few results, various researchers were found: “Stephen W Bell” as a collaborator of one article, “Stephen Bell” and “Stephen D Bell” from Indiana (likely both the same researcher, just under different aliases), and a “Stephen Bell” from London. Editing the search bar input to “Stephen P Bell” returns search results for articles containing the researcher by that name. At first glance, many of the articles do have affiliations to MIT. However, after further inspection into each of each article, some articles with “Stephen P Bell” are affiliated with Vermont. In such, the search bar input needs to be edited again to include MIT: “Stephen P Bell” “Massachusetts Institute of Technology”. Furthermore, if articles after a certain year (for example, 2010) were desired from these search results, a slider in the side bar must be used, since inputting a year into the search bar returns only for one year and inputting a range simply does not work for this database. Not only that, using “Stephen P Bell” with “Massachusetts Institute of Technology” or “MIT” provided different amounts of search results, despite being the same institution.

Next, after navigating to Google Scholar and following similar steps to before, searching for “Stephen Bell” returns many different people, including “Stephen Graham Bell” from the University of Adelaide—who is not the desired person in question. Changing the input to “Stephen P Bell” still results in variations of different people: one from MIT, one from the University of California, Berkeley, and one from Cold Spring Harbor Laboratory. While this may seem like an issue, upon further investigation, Stephen P Bell had attended all these institutions. In such, now “Stephen P Bell” “Massachusetts

Institute of Technology", "Stephen P Bell" "University of California, Berkeley", and "Stephen P Bell" "Cold Spring Harbor Laboratory" can all be used to search for the same individual in Google Scholar. Also, like PubMed, to filter the searches by year, a range must be set in the side bar, rather than inputting a value in the search bar, and using "Massachusetts Institute of Technology" returned a different number of results than "MIT".

NIH RePORTER has a similar issue with searching for "Bell, Stephen" and "Bell, Stephen P" as it returns other different people with similar, but different names. In such, "Stephen P Bell" "Massachusetts Institute of Technology" must be used ("University of California, Berkeley" returns no results and "MIT" returns less results).

Finally, there is the USPTO database. Navigating to the patent search is a little more tricky, but can be achieved by going to their main website, scrolling to "Patents" link near the bottom, then "Search patents" link under "Patent basics", then the link for "Patent Public Search" (twice if you click on it in the table of contents), then "Basic Search". Under the Basic search fields, change both "Search" to "Applicant Name", input "Bell" and "Stephen" in the "For" fields, and ensure the operator is "AND". Searching through all the search results reveals there is no Stephen P Bell.

Not only that, each one of the databases handles phrasing (surrounding a series of words with "double quotes" to indicate they are connected, should be searched together, and used to make more specific searches) differently. For example, PubMed treated "Stephen P Bell" (with quotes) returned the same number of results as that without quotes. This can confirm this by clicking the Advanced link under the search bar, and clicking through the search history, where it shows the entire name is grouped together. However, it can also become clear that this database doesn't treat "Massachusetts Institute of Technology" the same, as it separates each word for the searches (so using quotes for MIT would've been useful). Google Scholar returns more searches if the name is not within the quotations, so there is a benefit to using it in Google Scholar. NIH RePORTER still returned "Eugene Bell" with "Stephen P Bell"

in quotes, so other identifiers should be used (like MIT). Finally, USPTO treats words with quotations the same as those without due to already limiting one word per search box.

DISCUSSION

This research serves to add to this field of study and establish a general methodology to the overall matching algorithm, with a unique twist of utilizing synonyms and identifiers. These initial and secondary identifiers and synonym vocabularies help automatically and accurately match the information (essentially, attempting to replicate and automate the process of a person manually matching information between databases; after all, people identify matching data by detecting similarities in data).

Data matching allows for the compilation of information from multiple, different databases. When the data from each database is combined, it creates a large, integrated dataset that is useful for further analysis. Developing this general methodology will allow for improved assessments despite heterogenous databases, and more generally, provide a foundation for future studies to use—creating more research opportunities, collectively furthering progress in the field, and enhancing the detail of algorithm frameworks that will be developed—so that this issue may be eliminated in the future.

CONCLUSION

Due to this methodology being more generalized than preliminary research in the field, it can be helpful and applicable to multiple databases, fields, or simply serve as a framework for future research to be based off. That said, it is limited in the experience of the author, time spent in development, number of researchers, and does not have a specific matching algorithm—which could prevent the research from being immediately actionable. Subsequent research should increase the number of

researchers, the experience of the researchers, and the amount of time spent researching (this study took approximately a year in total, at a weekly pace, excluding portions of the summer semester, from late-May to early-August). Furthermore, due to the flexible nature of this framework, other matching algorithms could be supplemented (in place of the current algorithm) to improve the matching efficacy. Likewise, this study focuses on the algorithm of matching and combining data from databases, rather than delving into the research of research productivity. Following researchers should use this computational methodology to investigate research productivity more deeply. Not only that, a "master identifier" could be implemented, which would serve to designate the identifier that is more accurate (e.g., for some researchers, this may be their full name or NIH contact PI ID).

REFERENCES

Christen, P. (2014, February) Data Matching Research at the Australian National University.

de Leeuw, T., & Keijl, S. (2015). Research with Secondary Data: Different Matching Methods and is there a Difference?. In Academy of Management Proceedings (Vol. 2015, No. 1, p. 14215). Briarcliff Manor, NY 10510: Academy of Management.

Google. (n.d.). About Google Scholar. Google. Retrieved February 27, 2023, from https://scholar.google.com/intl/en/scholar/about.html.

Halpin, H., & McNeill, F. (2013). Discovering meaning on the go in large heterogenous data. Artificial Intelligence Review, 40, 107-126.

Kumar, A., Srivastava, R. P., Jadhav, G., Chaudhary, J., Kumar, A., Tyagi, M., Kumar, M., & Thakur, R. (Eds.). (2021, September 15). Linear Search. GeeksforGeeks. Retrieved April 25, 2022, from https://www.geeksforgeeks.org/linear-search/.

Landhuis, E. (2016). Scientific literature: Information overload. Nature, 535(7612), 457-458.

Lara, J. M. G., Osma, B. G., & Noguer, B. G. D. A. (2006). Effects of database choice on international accounting research. Abacus, 42(3-4), 426-454.

Merriam-Webster. (n.d.). Root. In Merriam-Webster.com dictionary. Retrieved February 16, 2023, from https://www.merriam-webster.com/dictionary/root.

National Center for Biotechnology Information (NCBI). (1996). PubMed Overview. U.S. National Library of Medicine (NLM). Retrieved February 26, 2023, from https://pubmed.ncbi.nlm.nih.gov/about/.

National Institutes of Health (NIH). (n.d.). About Us. NIH RePORT Research Portfolio Online Reporting Tools. Retrieved February 27, 2023, from https://report.nih.gov/about.

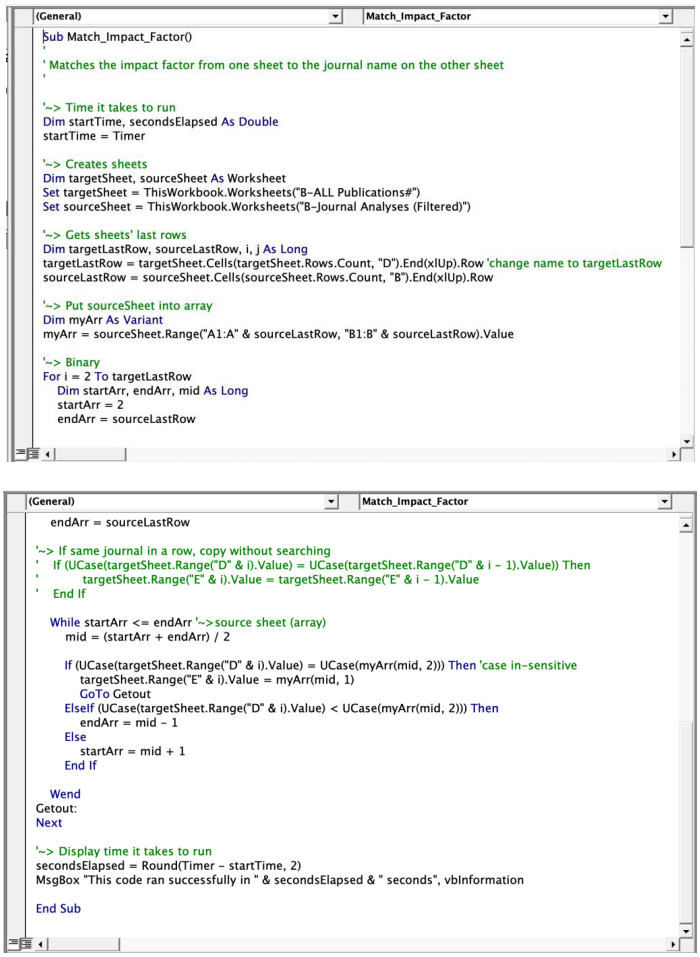
Pang, C., Gu, L., Hansen, D., & Maeder, A. (2009). Privacy-preserving fuzzy matching using a public reference table. Intelligent Patient Management, 71-89.

Srivastava, R. P., Jain, D., Kumar, P., Jha, A. K., Dwivedi, H. K., Arora, S., Soda, K., David, G., Kumar, G., & Ambati, K. (Eds.). (2022, March 25). Binary search. GeeksforGeeks. Retrieved April 25, 2022, from https://www.geeksforgeeks.org/binary-search/.

USPTO. (2018, June 5). About Us. USPTO United States Patent and Trademark Office. Retrieved February 27, 2023, from https://www.uspto.gov/about-us.

Appendix

Initially, this thesis sought to create a more in-depth computer program using Excel's programming language, Visual Basic for Applications (VBA). However, as it became evident that the primary researcher had inadequate experience in programming or the capability to create such a program, the project gradually developed into a general matching framework instead.



```
(General)                                   Match_Impact_Factor
Sub Match_Impact_Factor()
'
' Matches the impact factor from one sheet to the journal name on the other sheet
'

'~> Time it takes to run
Dim startTime, secondsElapsed As Double
startTime = Timer

'~> Creates sheets
Dim targetSheet, sourceSheet As Worksheet
Set targetSheet = ThisWorkbook.Worksheets("B-ALL Publications#")
Set sourceSheet = ThisWorkbook.Worksheets("B-Journal Analyses (Filtered)")

'~> Gets sheets' last rows
Dim targetLastRow, sourceLastRow, i, j As Long
targetLastRow = targetSheet.Cells(targetSheet.Rows.Count, "D").End(xlUp).Row 'change name to targetLastRow
sourceLastRow = sourceSheet.Cells(sourceSheet.Rows.Count, "B").End(xlUp).Row

'~> Put sourceSheet into array
Dim myArr As Variant
myArr = sourceSheet.Range("A1:A" & sourceLastRow, "B1:B" & sourceLastRow).Value

'~> Binary
For i = 2 To targetLastRow
   Dim startArr, endArr, mid As Long
   startArr = 2
   endArr = sourceLastRow
```

```
(General)                                   Match_Impact_Factor
   endArr = sourceLastRow

'~> If same journal in a row, copy without searching
'   If (UCase(targetSheet.Range("D" & i).Value) = UCase(targetSheet.Range("D" & i - 1).Value)) Then
'         targetSheet.Range("E" & i).Value = targetSheet.Range("E" & i - 1).Value
'   End If

   While startArr <= endArr '~>source sheet (array)
      mid = (startArr + endArr) / 2

      If (UCase(targetSheet.Range("D" & i).Value) = UCase(myArr(mid, 2))) Then 'case in-sensitive
         targetSheet.Range("E" & i).Value = myArr(mid, 1)
         GoTo Getout
      ElseIf (UCase(targetSheet.Range("D" & i).Value) < UCase(myArr(mid, 2))) Then
         endArr = mid - 1
      Else
         startArr = mid + 1
      End If

   Wend
Getout:
Next

'~> Display time it takes to run
secondsElapsed = Round(Timer - startTime, 2)
MsgBox "This code ran successfully in " & secondsElapsed & " seconds", vbInformation

End Sub
```

Figure 4. The VBA matching algorithm

The program (see Figure 4) was planned to search websites, search for a unit of analysis, exported into an Excel file, then matched together using a matching algorithm. To start, a linear search (see Figure 5) was used to match the data between Excel sheets but due to efficiency issues, a binary search algorithm was adopted instead (see Figure 6).

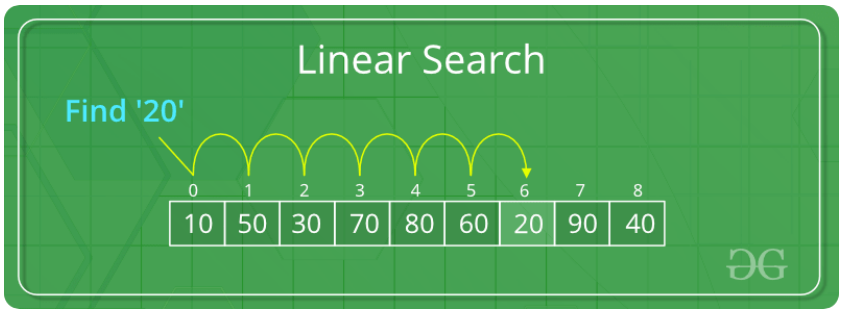


Figure 5. Linear search algorithm (Kumar, 2021)

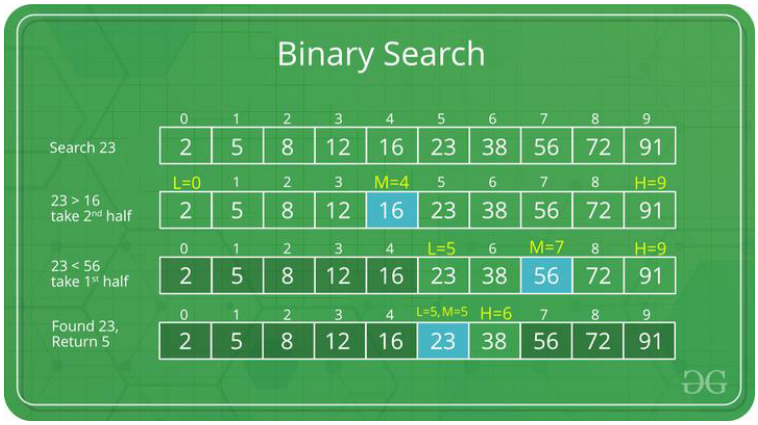
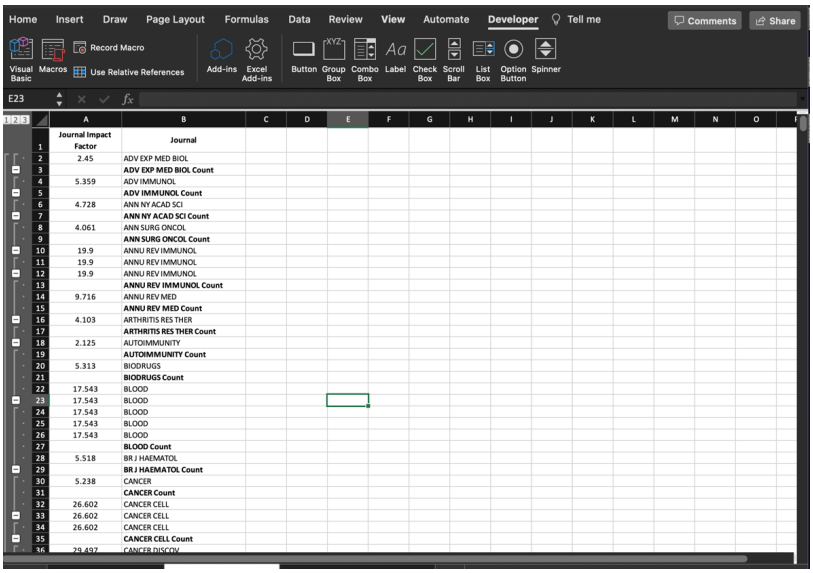


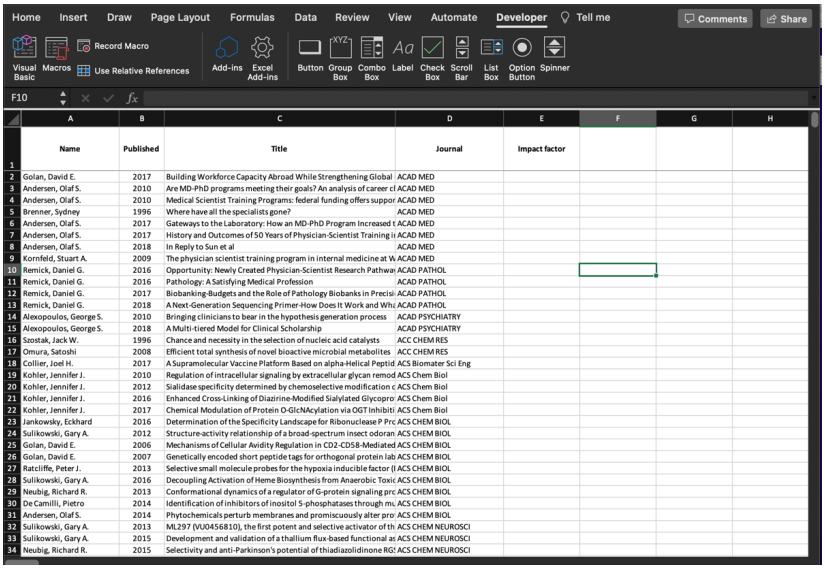
Figure 6. Binary search algorithm (Srivastava, 2022)

Specifically, the performance of the linear search took around 10 minutes on the Excel dataset used in this example. In contrast, the binary search algorithm took around 10 seconds (see Figure 9). The matching algorithm would match the impact factor in one Excel sheet (see Figure 7) to another Excel sheet (see Figure 8)—which would be the “integrated dataset” (see Figure 9).



	A	B
1	Journal Impact Factor	Journal
2	2.45	ADV EXP MED BIOL
3		**ADV EXP MED BIOL Count**
4	5.359	ADV IMMUNOL
5		**ADV IMMUNOL Count**
6	4.728	ANN NY ACAD SCI
7		**ANN NY ACAD SCI Count**
8	4.061	ANN SURG ONCOL
9		**ANN SURG ONCOL Count**
10	19.9	ANNU REV IMMUNOL
11	19.9	ANNU REV IMMUNOL
12	19.9	ANNU REV IMMUNOL
13		**ANNU REV IMMUNOL Count**
14	9.716	ANNU REV MED
15		**ANNU REV MED Count**
16	4.103	ARTHRITIS RES THER
17		**ARTHRITIS RES THER Count**
18	2.125	AUTOIMMUNITY
19		**AUTOIMMUNITY Count**
20	5.313	BIODRUGS
21		**BIODRUGS Count**
22	17.543	BLOOD
23	17.543	BLOOD
24	17.543	BLOOD
25	17.543	BLOOD
26	17.543	BLOOD
27		**BLOOD Count**
28	5.518	BR J HAEMATOL
29		**BR J HAEMATOL Count**
30	5.238	CANCER
31		**CANCER Count**
32	26.602	CANCER CELL
33	26.602	CANCER CELL
34	26.602	CANCER CELL
35		**CANCER CELL Count**
36	29.497	CANCER DISCOV

Figure 7. The Excel sheet with the impact factor



	A	B	C	D	E	F	G	H
1	Name	Published	Title	Journal	Impact factor			
2	Golan, David E.	2017	Building Workforce Capacity Abroad While Strengthening Global	ACAD MED				
3	Andersen, Olaf S.	2010	Are MD-PhD programs meeting their goals? An analysis of career cl	ACAD MED				
4	Andersen, Olaf S.	2010	Medical Scientist Training Programs: federal funding offers suppor	ACAD MED				
5	Brenner, Sydney	1996	Where have all the specialists gone?	ACAD MED				
6	Andersen, Olaf S.	2017	Gateways to the Laboratory: How an MD-PhD Program Increased t	ACAD MED				
7	Andersen, Olaf S.	2017	History and Outcomes of 50 Years of Physician-Scientist Training i	ACAD MED				
8	Andersen, Olaf S.	2018	In Reply to Sun et al	ACAD MED				
9	Kornfeld, Stuart A.	2009	The physician scientist training program in internal medicine at W	ACAD MED				
10	Remick, Daniel G.	2016	Opportunity: Newly Created Physician-Scientist Research Pathwa	ACAD PATHOL				
11	Remick, Daniel G.	2016	Pathology: A Satisfying Medical Profession	ACAD PATHOL				
12	Remick, Daniel G.	2017	Biobanking-Budgets and the Role of Pathology Biobanks in Precisi	ACAD PATHOL				
13	Remick, Daniel G.	2018	A Next-Generation Sequencing Primer-How Does It Work and Wha	ACAD PATHOL				
14	Alexopoulos, George S.	2010	Bringing clinicians to bear in the hypothesis generation process	ACAD PSYCHIATRY				
15	Alexopoulos, George S.	2018	A Multi-tiered Model for Clinical Scholarship	ACAD PSYCHIATRY				
16	Szostak, Jack W.	1996	Chance and necessity in the selection of nucleic acid catalysts	ACC CHEM RES				
17	Omura, Satoshi	2008	Efficient total synthesis of novel bioactive microbial metabolites	ACC CHEM RES				
18	Collier, Joel H.	2017	A Supramolecular Vaccine Platform Based on alpha-Helical Peptid	ACS Biomater Sci Eng				
19	Kohler, Jennifer J.	2010	Regulation of intracellular signaling by extracellular glycan remod	ACS Chem Biol				
20	Kohler, Jennifer J.	2012	Sialidase specificity determined by chemoselective modification c	ACS Chem Biol				
21	Kohler, Jennifer J.	2016	Enhanced Cross-Linking of Diazirine-Modified Sialylated Glycopro	ACS Chem Biol				
22	Kohler, Jennifer J.	2017	Chemical Modulation of Protein O-GlcNAcylation via OGT Inhibiti	ACS Chem Biol				
23	Jankowsky, Eckhard	2016	Determination of the Specificity Landscape for Ribonuclease P Prc	ACS CHEM BIOL				
24	Sulikowski, Gary A.	2012	Structure-activity relationship of a broad-spectrum insect odoran	ACS CHEM BIOL				
25	Golan, David E.	2006	Mechanisms of Cellular Avidity Regulation in CD2-CD58-Mediated	ACS CHEM BIOL				
26	Golan, David E.	2007	Genetically encoded short peptide tags for orthogonal protein lab	ACS CHEM BIOL				
27	Ratcliffe, Peter J.	2013	Selective small molecule probes for the hypoxia inducible factor (I	ACS CHEM BIOL				
28	Sulikowski, Gary A.	2016	Decoupling Activation of Heme Biosynthesis from Anaerobic Toxic	ACS CHEM BIOL				
29	Neubig, Richard R.	2013	Conformational dynamics of a regulator of G-protein signaling pro	ACS CHEM BIOL				
30	De Camilli, Pietro	2014	Identification of inhibitors of inositol 5-phosphatases through mu	ACS CHEM BIOL				
31	Andersen, Olaf S.	2014	Phytochemicals perturb membranes and promiscuously alter pro	ACS CHEM BIOL				
32	Sulikowski, Gary A.	2013	ML297 (VU0456810), the first potent and selective activator of th	ACS CHEM NEUROSCI				
33	Sulikowski, Gary A.	2015	Development and validation of a thallium flux-based functional as	ACS CHEM NEUROSCI				
34	Neubig, Richard R.	2015	Selectivity and anti-Parkinson's potential of thiadiazolidinone RG	ACS CHEM NEUROSCI				

Figure 8. The Excel sheet that the impact factor is copied to