Published in final edited form as:

Procedia Comput Sci. 2022 ; 211: 196–200. doi:10.1016/j.procs.2022.10.191.

Data Integration for the Study of Outstanding Productivity in Biomedical Research

Clément Aubert[a,*], **E Andrew Balas**[a], **Tiffany Townsend**[a], **Noah Sleeper**[a], **C.J. Tran**[a]

[a]Augusta University, GA, USA

Abstract

Our goal is to analyze improvement of scientific performance in a multidimensional outcome space, with a focus on US-based biomedical research. With the growing diversity of research databases, limiting assessment of scientific productivity to bibliometric measures such as number of publications, impact factor of journals and number of citations, is increasingly challenged. Using a wider range of outcomes, from publications through practice improvements to entrepreneurial outcomes, overcomes many current limitations in the study of research growth. However, combining such heterogeneous datasets raise three challenges: 1. gathering in one common place a variety of data shared as csv, xml or xls files, 2. merging and linking this data, that sometimes overlap, 3. assessing the impact of research production and inclusive practices in a multidimensional space, that are often missing from the datasets. We would like to present our solution for the first of those challenges, and discuss our leads for the second and third challenges.

Keywords

Biomedical Research; Scientific Performance; Matching of Research Databases; Research Evaluation

1. Introduction – Need for Multidimensional Assessment Of Research Performance

Previously, bibliometric and publication databases have been the most advanced sources of information on research performance. More recently, citation statistics or download counts have become widely available. With the growing diversity of research databases, limiting assessment of scientific productivity to bibliometric measures such as number of publications, impact factor of journals and number of citations is increasingly challenged [8]. Furthermore, research assessment based on publications in international journals has become insufficient in the eyes of policy makers [5].

Among other available indicators are federal research funding received, completed clinical trials, patents issued, or start-up companies creations, to name a few. Those can help in driving a multi-faceted approach that could make “[r]esearch quality improvement […]

*Corresponding author. *URL:* https://spots.augusta.edu/caubert/ (Clément Aubert), caubert@augusta.edu.

a continuous and comprehensive process, from the design and conduct of research to the publication of results." [10] Adopting comprehensive measures not based solely on bibliometrics factors will help in reducing deficiencies, increasing research productivity, and enabling meritorious scientific discoveries.

However, studying the outcomes of research is essential for determining whether the society's research investments are paying off, but the abstract focus on such outputs may diminish its quality [3]. A study of academic medical centers highlighted that research core facilities and platforms are often evaluated only for putting out fires, like continued annual deficits, instead of aligning and evaluating them strategically [7].

Even if improving the quality and reproducibility of research is a major societal interest, there is a scarcity of studies on biomedical research growth strategies. Growth in research universities or institutions is much more widely discussed, but the discussions are rarely data driven [2]. Occasionally, international university rankings are used as goal setters for research growth with limited effectiveness [12].

Our goal is to analyze improvement of scientific performance in a multidimensional outcome space, with a focus on biomedical research in the United States of America. Our contribution will

1. provide an excellent case study,
2. sketch interesting operational solutions to data agglomeration,
3. develop innovative ways of assessing research productivity, and
4. explore the influence of laboratory diversity and inclusive practices on the quality of biomedical research.

The problems we are facing are not new. As a recent study on data integration puts it, "two important challenges for the field [… are …] to develop good open-source tools for different components of data integration pipelines [… and …] to provide practitioners with viable solutions for the long-standing problem of systematically combining structured and unstructured data." [6] While we believe our approach partially resolves the first challenge, performing high-quality, traceable, linking between our numerous datasets prove to be a great challenge.

2. Sketch of the Proposed Approach

The proposed research gathers a multidisciplinary team composed of a researcher in Computer Science, a Dr. in interdisciplinary health sciences, a chief diversity officer/health equity researcher and two students to address this problem originally. In a nutshell, our approach intends to leverage a wide range of emerging scientific databases to isolate, study and compare research institutions with outstanding productivity. The proposed study will also investigate the driving factors, like workforce and diversity, behind outstanding productivity, even if those may be harder to extract from available data: if necessary, targeted surveys, interviews, panel discussions as well as other ingenious methodologies—

i.e., comparing the authors' names with U.S. Census data to obtain their ethnic identification [1]—will be required to fill this gap and create the required data.

The variety of emerging scientific databases and their rapidly improving access opportunities make possible more multifaceted and granular study of progress in research. We believe that the use of a wide range of outcomes, from publications through practice improvements to entrepreneurial outcomes, overcomes many current limitations in the study of research growth. The expertise and competency analyses will create opportunities to identify and increase the value of collaborative research that places priority on diversity of the team and interdisciplinary research. The proposed study is innovative because it significantly expands the performance analysis of the biomedical research enterprise by identifying and testing a large variety of new metrics based on the rapidly growing selection of pertinent databases (cf. Table 1). Unfortunately, those datasets use extremely diverse representations and format (`csv, xml, xls` files that do not use common schemas), and they do not use persistent identifiers for researchers or organizations, as each uses their own identification mechanism (which sometimes consists only of a name and email address). This limits the ability to use e.g., research graphs and other readily available solutions to cross-linking these entities.

Due to the inherent diversity of those data sources, our first challenge is to curate those datasets. To this end, a significant effort has been devoted to

1. identify relevant attributes in the datasets,
2. harvest the data as `csv, xml, xls`, or even sometimes `html` files[1],
3. insert them in a SQL database,
4. output them to a `xls` file to ease mathematical treatment and statistical analysis by the team.

Each dataset comes with a precise, documented schema, that can be leveraged, if needed using genetic algorithms [11], manual annotations [14], or semantics constraints [9] to insure the best possible quality of the extracted data. A fully operational prototype written in Java and available on-line [13] can perform those tasks (cf. Figure 1, Steps 1 and 3) but remains the delicate problem of matching (or linking [4, 16, 17]) those various sources that have wildly different schemas and organizational practices (cf. Figure 1, Step 2).

Typically, a unique researcher, laboratory or institution can be identified differently in any two databases (combining first and last name into a single field, using alternate spelling or abbreviations), not to mention possible change of name or affiliation or spelling mistakes. Unfortunately, standard such as the Common European Research Information Format (CERIF) [15] do not exist in the US, or at least are not used in the targeted databases. Hence, combining this data requires a reliable integration methods that is mostly automated but supportive of visual error checking as well. Our multi-tool platform allows to address

[1]This particular effort is led by leveraging python's https://www.selenium.dev/ API, and integrating its harvested data into our pipeline. We hope to share this effort publicly as well soon.

this problem with SQL and Excel tools, but even so ensuring a good quality, traceable and accountable linking remains a challenge.

While our model area is restricted to the biomedical research, it is our hope that our tools and findings will be widely re-usable and of interest to many different communities. Conversely, we hope to draw on the CRIS systems that collect data on entrepreneurial outcomes, clinical trials, competitive research grants received to rapidly adopt best practises and build on their expertise and tools.

Acknowledgements

This work was supported by the grant R01 GM146338 from the NIH National Institute of General Medical Sciences in the SCISIPBIO program. The authors would like to thank the reviewers for their interesting comments, that greatly improved our presentation.

References

[1]. Beardsley R, Halevi G, 2022. Insights: Ethnic diversity in STEM in the United States. Technical Report. Institute for Scientific Information. doi:10.14322/isi.insight.1.

[2]. Birx DL, Anderson-Fletcher E, Whitney E, 2013. Growing an emerging research university. Journal of Research Administration 44, 11–35. URL: https://eric.ed.gov/?id=EJ1013309.

[3]. Bowen A, Casadevall A, 2015. Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. Proceedings of the National Academy of Sciences 112, 11335–11340. doi:10.1073/pnas.1504955112.

[4]. Elmagarmid AK, Ipeirotis PG, Verykios VS, 2007. Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19, 1–16. doi:10.1109/TKDE.2007.250581.

[5]. Ernø-Kjølhede E, Hansson F, 2011. Measuring research performance during a changing relationship between science and society. Research Evaluation 20, 131–143. doi:10.3152/095820211X12941371876544, arXiv:https://academic.oup.com/rev/article-pdf/20/2/131/4484270/20-2-131.pdf.

[6]. Golshan B, Halevy A, Mihaila G, Tan WC, 2017. Data integration, in: Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, ACM. pp. 101–106. doi:10.1145/3034786.3056124.

[7]. Haley R, Champagne TJ Jr, 2017. Research strategies for academic medical centers: A framework for advancements toward translational excellence. Research Management Review 22, n1. URL: https://eric.ed.gov/?id=EJ1134104.

[8]. Lindner MD, Torralba KD, Khan NA, 2018. Scientific productivity: An exploratory study of metrics and incentives. PLOS ONE 13, 1–16. doi:10.1371/journal.pone.0195321.

[9]. Lv T, Yan P, 2006. Mapping dtds to relational schemas with semantic constraints. Information and Software Technology 48, 245–252. URL: http://www.sciencedirect.com/science/article/pii/S0950584905000777, doi:10.1016/j.infsof.2005.05.001.

[10]. Mansour NM, Balas EA, Yang FM, Vernon MM, 2020. Prevalence and prevention of reproducibility deficiencies in life sciences research: Large-scale meta-analyses. Medical Science Monitor 26. URL: 10.12659/msm.922016, doi:10.12659/msm.922016.

[11]. Ng V, Kong CC, Chan S, 2004. Mapping xml schema to relations using genetic algorithm, in: Negoita MG, Howlett RJ, Jain LC (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 232–245. doi:10.1007/978-3-540-30134-9_33.

[12]. Sitnicki MW, 2018. Determining the priorities of the development of eu research universities based on the analysis of rating indicators of world-class universities. TalTech Journal of European Studies 8, 76–100. URL: 10.1515/bjes-2018-0006, doi:doi:10.1515/bjes-2018-0006.

[13]. Sleeper N, Aubert C, 2022. Data Integration for the Study of Outstanding Productivity in Biomedical Research. URL: https://github.com/popbr/data-integration.

[14]. Sunchu VK, 2016. A Flexible Schema-Aware Mapping of XML Data into Relational Models. Master's thesis. The University of Oklahoma, College of Engineering, School of Computer Science. URL: https://hdl.handle.net/11244/44903.

[15]. The euroCRIS CERIF Task Group, 2021. CERIF-DataModel. URL: https://github.com/EuroCRIS/CERIF-DataModel.

[16]. Winkler WE, 1999. The State of Record Linkage and Current Research Problems. Technical Report. Statistical Research Division, U.S. Census Bureau.

[17]. Winkler WE, 2006. Overview of record linkage and current research directions. Technical Report. Statistical Research Division, U.S. Census Bureau.

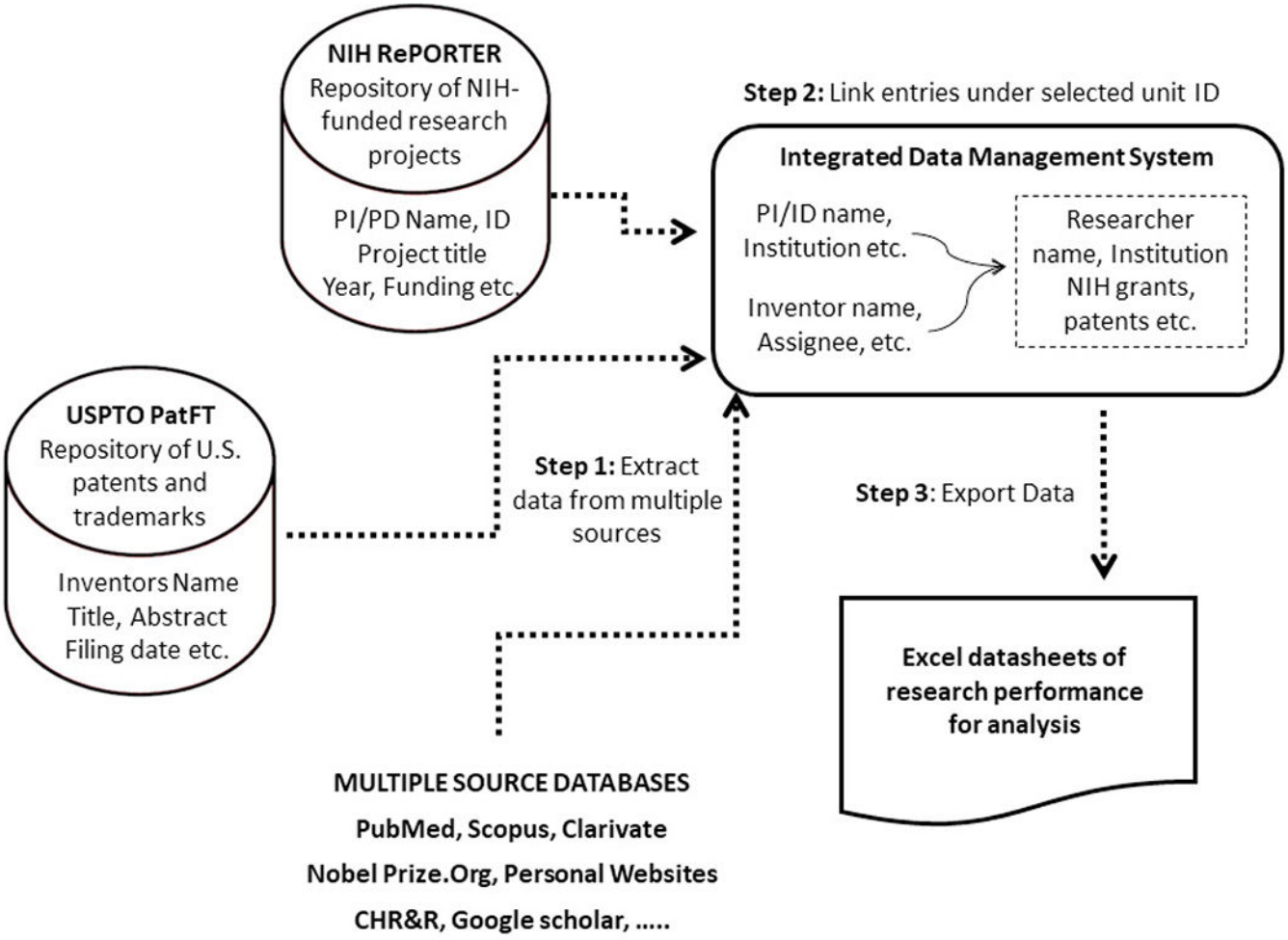

Fig. 1.
Collecting, Linking and Exporting Heterogeneous Datasets

Table 1.

Illustrative indicators of the three-dimensional value of scientific research

Dimension	Indicator	Description	Source
Scientific	Publications	Average peer-reviewed publications	PubMed Database NIH, National Library of Medicine, WoS, RePORT via ExPORTER
	Citations	Average publications in the top 10% from WoS Core Collection	WoS
	Competitive research grants received	Average federal research expenditures received	NSF Award Search, RePORT via ExPORTER
Public health	Completed clinical trials	Average completed clinical trials completed	Clinicaltrials.gov
	Contributions to FDA approved products	Average patents associated with institution and FDA device or drug approval	FDA Orange Book, PATSTAT, WoS
	Contributions to clinical practice guidelines (CPG)	Average publications cited within a published CPG	AHRQ CPG Clearinghouse, PubMed, WoS
Economic	Joint publications with industry	Average publications jointly published with industry	WoS
	Startups	Average number of start-up companies founded by institution	AUTM Survey, D&B Hoovers database, STATT
	Gross Licensing Income	Average gross income received from IP licensing	AUTM Survey
	Social and economic effects of research investments	Average expenditure received by institution	IRIS Data
	Patents and Trademarks	Average number of patents and trademarks issued by institution	USPTO
	Investment	Research and development expenditures	NSF HERD
Multi-dimensional	-	-	Dimensions