Data Integration for the Study of Outstanding Productivity in Biomedical Research

C. Aubert, A. Balas, T. Townsend, N. Sleeper and C.J. Tran (Augusta University, USA)

March 11, 2022

Introduction – Imperiosity of Assessing Scientific Performances

Our goal is to analyze improvement of scientific performance in a multidimensional outcome space, with a focus on American biomedical research. Even if the Current Research Information Systems privileges data and systems across Europe, we believe our contribution will a) provide an excellent case study, b) sketch interesting operational solutions to data agglomeration, c) develop innovative ways of assessing research productivity, and d) explore the influence of laboratory diversity and inclusive practices on the quality of biomedical research.

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 (Lindner 2018). Furthermore, research assessment based on publications in international journals has become insufficient in the eyes of policy makers (Ernø-Kjølhede and Hansson 2011).

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 (Bowen and Casadevall 2015). 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 (Haley and Champagne Jr 2017).

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 (Birx, Anderson-Fletcher, and Whitney 2013). Occasionally, international university rankings are used as goal setters for research growth with limited effectiveness (Sitnicki 2018).

Sketch of the Proposed Approach

The proposed research gathers an interdisciplinary team composed of a a researcher in Computer Science, a Dr. in interdisciplinary health sciences, a chief diversity officer/health equity researcher and two Computer Science 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 and panel discussions will 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).

Due to the inherent diversity of those data sources, however, a significant effort is devoted to a) harvest the data as `csv`, `xml`, `xls`, or even sometimes `html` files, b) insert them in a SQL database, c) output them to a `xls` file to ease mathematical treatment and statistical analysis by the team (cf. Figure 1). A fully operational prototype written in Java can perform those three tasks but remains the delicate problem of matching (or linking) those various sources that have wildly different schemas and organizational practices.

Typically, a unique researcher, laboratory or institution can be identified differently in any two database (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. 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 this problem with SQL and Excel tools, but even so ensuring a good quality, traceable and accountable linking remains a challenge, especially since some datasets may overlap (cf. Figure 2). It is our hope that the 15th International Conference on Current Research Information Systems will be at the same time interested in our original problem, in our innovative approach, but also help us shape a re-usable and pertinent solution to our linking problem.

This work was supported by the grant R01 GM146338 from the NIH National Institute of General Medical Sciences in the SCISIPBIO program.

Appendix

Table 1: Illustrative indicators of the three-dimensional value of scientific research

Dimension	Indicator	Description	Source
Scientific	Publications	Average peer-reviewed publications	WoS, NIH Research Portfolio Online Reporting Tools (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 Higher Education Research and Development Survey (HERD), NIH Research Portfolio Online Reporting Tools (RePORT) via ExPORTER
Public health	Completed clinical trials	Average completed clinical trials completed 2011 - 2015	Clinicaltrials.gov
	Contributions to FDA approved products	Average patents associated with institution and FDA device or drug approval	FDA Orange Book, PATSTAT, WoS

Dimension	Indicator	Description	Source
	Contributions to clinical practice guidelines (CPG)	Average publications cited within a CPG published in 2014	AHRQ Clinical Practice Guidelines 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	ATUM Annual Licensing Survey 2011-2015, D&B Hoovers database of in-depth Company Profiles, STATT: Statistics Access for Technology Transfer Database
	Gross Licensing Income	Average Gross income received from IP licensing	ATUM Annual Licensing Survey 2011-2015

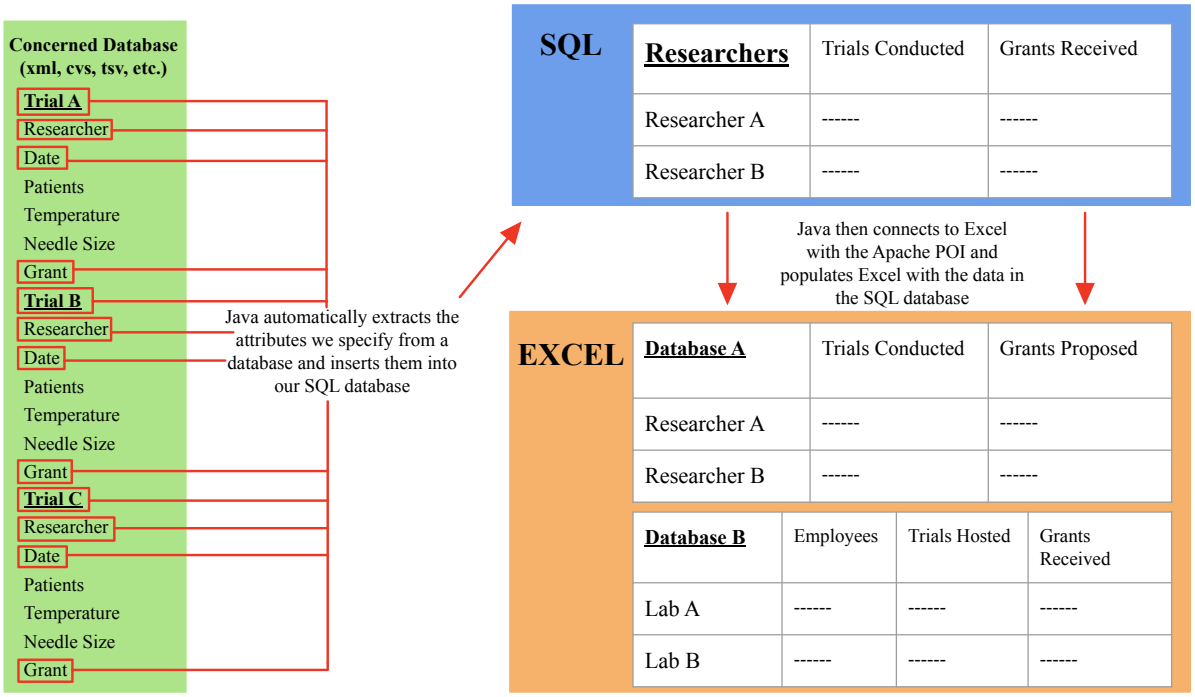

Figure 1: Merging Datasets Into one Table

References

Birx, Donald L, Elizabeth Anderson-Fletcher, and Elizabeth Whitney. 2013. "Growing an Emerging Research University." *Journal of Research Administration* 44 (1): 11–35. https://eric.ed.gov/?id=EJ1013309.

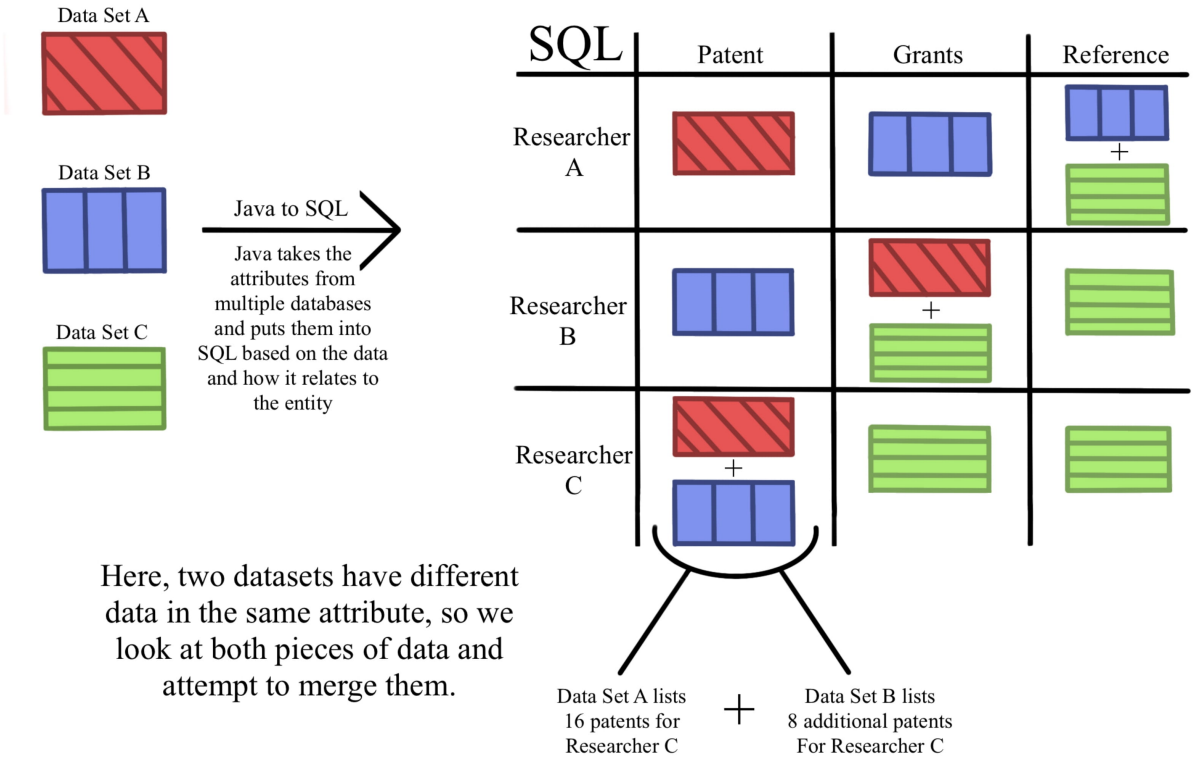

Figure 2: Merging and Linking Redundant Datasets

Bowen, Anthony, and Arturo Casadevall. 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 (36): 11335–40. https://doi.org/10.1073/pnas.1504955112.

Ernø-Kjølhede, Erik, and Finn Hansson. 2011. "Measuring research performance during a changing relationship between science and society." *Research Evaluation* 20 (2): 131–43. https://doi.org/10.3152/095820211X12941371876544.

Haley, Rand, and Thomas J Champagne Jr. 2017. "Research Strategies for Academic Medical Centers: A Framework for Advancements Toward Translational Excellence." *Research Management Review* 22 (1): n1. https://eric.ed.gov/?id=EJ1134104.

Lindner, Karina D. AND Khan, Mark D. AND Torralba. 2018. "Scientific Productivity: An Exploratory Study of Metrics and Incentives." *PLOS ONE* 13 (4): 1–16. https://doi.org/10.1371/journal.pone.0195321.

Sitnicki, Maksym W. 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 (1): 76–100. https://doi.org/doi:10.1515/bjes-2018-0006.