Scientific Findings as Big Data for Research Synthesis: The metaBUS Project

Frank Bosco
School of Business
Virginia Commonwealth University
Richmond, VA
fabosco@vcu.edu

Krista Uggerslev
JR Shaw School of Business
Northern Alberta Institute of Technology
Edmonton, AB
kristau@nait.ca

Piers Steel
Haskayne School of Business
University of Calgary
Calgary, AB
piers.steel@haskayne.ucalgary.ca

***Abstract*—We describe the metaBUS project, a large-scale research curation effort supported by the Digging into Data Challenge. This ongoing effort involves the extraction and curation of a corpus of more than one million individual research findings from organizational research. The approach involves the development of a comprehensive hierarchical taxonomy containing thousands of variables studied in the scientific space. Technological enablements allow linkages between the taxonomy and research findings fostering the development of a research finding search engine at the level of individual primary studies that allows the conduct of instant meta-analyses on virtually any relation of interest in organizational research.**

***Keywords*-big data; research curation; meta-analysis**

I. Introduction

Researchers in the social and medical sciences have long acknowledged that findings from a single empirical study can rarely, if ever, settle scientific questions [1]. The reasons for this, particularly within the areas of social and medical science, include measurement error (i.e., unreliability), sampling error, low statistical power, and publication bias [2], among others. As one approach to at least partially ameliorating such concerns, quantitative research synthesis has emerged and increased in popularity over a relatively short time. In research synthesis, relevant scientific findings (e.g., effect sizes) are sought out, classified, and submitted to meta-analysis. Typically, the meta-analysts goal is to ascertain (a) the overall mean effect size for a given relation of interest, (b) its dispersion, and (c) moderating effects acting upon the relation. Importantly, published meta-analytic results may guide further primary research in the area and become some of the most highly cited works in the academic literature, and they are significantly more likely than results of individual primary studies to reach the general public (i.e., through news reports or textbooks) [3]. Thus, meta-analytic findings have the potential to play an increasingly central role in the advancement of research and public understanding of science.

However, systematic reviews are not necessarily easy to conduct and the process can take years from planning to publication. In addition, a decade or more can pass before a revised review is published [4]–[7]. Provided that the number of scientific journals and, thus, scientific findings, is increasing each year, the process of conducting a systematic review will only become more difficult and time-consuming. In addition, as scientific disciplines expand, one can also expect that literature reviews will increasingly reflect where science once was – not where we are now. In short, the steadily increasing flood of new data [8] will pose a greater and greater challenge for those conducting systematic reviews, perhaps even to the point where the low return on investment for research teams conducting such reviews will view the approach as relatively unpalatable for meeting the academic goal of publishing research.

One of the most difficult and time-consuming aspects of conducting systematic reviews involves searching the literature for relevant findings. As an example, [9] conducted an article-by-article search (i.e., by hand) of 378 journal volumes. Using a conservative estimate of 6 issues per volume and 8 articles per issue, [9] examined each of 18,144 published articles in their search for findings on the relation between employee job satisfaction and job performance, a daunting task. While this approach to identifying studies examining the relation is highly comprehensive, each relation must then be arduously coded along a number of aspects related to the item, variable, and study, which is a second difficult and time-consuming aspect of systematic reviews. And the large effort expenditure of identifying and then coding relations provides little return on investment to consumers of science beyond the single reported meta-analytic estimate of interest. In short, current search processes for quantitative reviews represent a monument of inefficiency – a Digging into Data challenge.

As a third challenge, the corpus of constructs in the social sciences appears to suffer from the so-called jingle-jangle problem which, similar to the vocabulary problem [10], refers to a phenomenon whereby empirically indistinguishable constructs take on a variety of names and the same name may be used for two distinct constructs. For example, [11] reported that two constructs studied by applied psychologists, job satisfaction and organizational commitment, are so highly related (correlation corrected for unreliability = .91) that it is highly unlikely that the two actually represent distinct constructs. One possible downstream effect of the proliferation of constructs (or, at least, construct names)

in some social sciences is an overcomplicated theoretical landscape [12] that only makes the task of summarizing and communicating our findings to the public more difficult.

To expand on the challenge described above, the construct proliferation problem has led to a necessary silo arrangement among scientists themselves because the current level of nuance is far from ideal given our lack of research curation. Are social scientists really in need of 20 or 30 constructs that all refer to an employee's level of satisfaction with his or her supervisor? Are these constructs actually distinguishable from one another? If not, then we are in dire need of some theory pruning [12]. If so, then we are also in need of some sort of roadmap – a science construct map that provides a comprehensive layout of our field and, perhaps, corresponding Rosetta Stones that describe those categories of constructs in lay terms. The metaBUS approach is able to handle both of these lingering concerns.

Applied psychologists are all too familiar with the scientist-practitioner gap, a problem characterized by a lack of scientific understanding on the part of (would-be) science consumers, such as human resource (HR) managers. According to a recent large-scale survey of HR managers, 72 percent agree with the false statement that tests of personality are superior employee selection devices compared with tests of cognitive ability [13]. However, the extant evidence supports the reverse - cognitive ability outperforms personality in most contexts by approximately a factor of three, and scientists have been aware of this general finding for decades [14]. Clearly, researchers in many of the social sciences are in dire need of approaches to curate and communicate their findings in ways that are readily digestible for practitioners to use.

The remainder of our manuscript is organized as follows. First, we describe existing approaches to science mapping. Next, we describe our ongoing Digging into Data Challenge project with the aims of (a) collecting all scientific findings in a specified scientific space, (b) categorizing each finding according to a comprehensive taxonomy of constructs, and (c) providing open access to scientific and practitioner audiences with automated empirical summaries. In short, we describe an approach for developing a platform with which to map scientific constructs, curate existing findings, and conduct instant meta-analyses from the universe of findings in social science fields.

A. Science Mapping

Maps are representations of complex spaces and are useful to the extent that they summarize a body of information; maps condense large amounts of information and, thus, facilitate navigation and understanding [15]. The usefulness of science maps is eloquently described by [16], who remind us of the Indian fable of the blind men and the elephant. According to the fable, six blind men attempt to identify the animal after having felt a different part of it. Not surprisingly, the blind men lack consensus in their identification of the animal. However, ascertaining the shape and nature of a scientific discipline as a whole is a much more complex task. Indeed, as described by [16], "'But science does not stand still; the steady stream of new scientific literature creates a continuously changing structure. The resulting disappearance, fusion, and emergence of research areas adds another twist to the tale it is as if the elephant is running and dynamically changing its shape"' (p. 180). Science mapping is thus highly valuable because it provides a big picture view of a field and allows a better understanding of its nature [17], [18].

Approaches to science mapping vary in terms of their scope and content [19]. Global science maps attempt to represent all scientific disciplines, providing a large-scale view of interrelationships between fields, whereas local science maps provide a detailed view of a particular discipline or subdiscipline [20]. Science maps extract meaning-based relationships by analyzing the co-occurrence of references or keywords. That is, articles are considered related to the extent that they share references, authors, or keywords. The approach has been applied in related literatures [21]–[23] and can be considered, broadly, as bibliographic approaches. However, existing science mapping approaches do relatively little to communicate the actual findings located in their varied science corpora. As an example, [24] located articles on the topic of careers by searching article titles for the letter string career. Importantly, approaches like these likely overlook a large portion of the related research corpus.

B. Research Archives

Several research finding archives exist across the sciences. As examples, the Cochrane Collaboration [25] archives and makes available existing meta-analytic evidence for medical research. A much smaller social science version is the Campbell Collaboration Library of Systematic Review [26]. Similarly, the Research Findings Register (ReFeR), administered by the UKs Department of Health, provides access to funded health research findings. The NRR archive at the National Institute for Health Research (UK) provides a similar database. Archives are available for genomic research [27] focusing on the management of incidental findings. In addition, archival efforts have been undertaken within focused areas of study, such as happiness research [28]. However, these approaches, while involving large databases of findings, do little to curate the findings in a way that makes navigation and summarization relatively easy.

Possibly the closest version to our endeavor is the recently developed Systematic Review Data Repository (SRDR), with support by the US Department of Healthcare Research and Quality. Though entirely purposed for healthcare, this resource would serve as both a central archive and data extraction tool, shared among and freely accessible to organizations producing systematic reviews worldwide (p. 15)

[29]. As [29] note, that while there are several similar endeavors in the medical field, such as research registries and commercial databases, "'this proposed archive is unique in its scope and mission, and will provide a tool sorely lacking in the research community"' (p. 20).

II. Methodologies

A. Overview

Our approach involves (1) extraction of correlation matrices from PDF files, (2) development of a comprehensive taxonomy of constructs, (3) migration of extracted matrices to a database, where they are processed, cleaned, and formatted for meta-analysis, and (4) addition of formatted data to a master database, where coders manually add article- and variable-level codes (e.g., taxonomy codes). An integrated software application currently under development will enable real-time meta-analyses on every topic imaginable within the field of human resources and organizational behavior by scientists and practitioners around the world.

B. Data Extraction

Our project involves archiving research findings located in published scientific journal articles. In many of the social sciences, research findings take the form of effect sizes (e.g., correlation coefficients) that are often arranged in correlation matrices in published journal articles. As shown in Figure 1 (lower panel), a correlation matrix represents an impressive repository of information. Importantly, the amount of information presented in salient article search text (Figure 1, upper panel) rarely reflects all of the information available in a given article, although each piece of information contained in the matrix is perfectly suitable for inclusion in research synthesis. In Figure 1, the values highlighted in green represent what one would expect to locate in terms of research findings, from exposure to the salient article text. The values highlighted in red (most of the data in the article) would have been located only by browsing the article. Typically, a matrix will present the names of all studied variables, their mean and standard deviation values, reliabilities along the matrixs diagonal axis, and correlations between all possible variable pairs. Extrapolating from the rate of our ongoing archival effort, we expect to accumulate approximately one million individual research findings from every article from 1990 through 2013 across 30 journals in the field.

Importantly, we expect that full-text article searches will be unable to efficiently capture all of the data located in the lower panel of Figure 1. To elaborate, imagine that a research team were conducting a systematic review on the relation between employee age and employee performance. Clearly, a full-text search of articles containing the letter strings age and performance, within organizational research outlets, would likely return virtually all articles in the field. Thus, full-text search is not a sufficient replacement for curation of findings. Natural language processing approaches could offer some degree of efficiency, however, we expect that a fully-automated solution that offers a high level of taxonomic coding accuracy is not possible at this time. Indeed, the construct proliferation problem described above only further complicates matters and makes a fully-automated solution less likely.

C. Development of Taxonomies

The development of taxonomies for any corpus of data is a difficult process. Indeed, as described by [30] regarding a similar classification challenge, "'classifying data into categories made description and navigation relatively easy at the same time, it named things and tried to fit the diversity of our world into predetermined buckets that reflected more the biases of the classifying entity rather than our reality"' (p. 34). Earlier approaches to classifying constructs in organizational research do so according to a construct's purpose (e.g., personality tests as predictors of employee performance [31]). However, reflecting on this example, personality tests are used in organizational research for predicting a plethora of outcomes, such as employee turnover intentions. Our approach involves classifying constructs according to what they are rather than how they are used, a much more general approach that is relatively defendable. Thus, to elaborate on the example, we classify personality traits under in our taxonomy as person characteristics - psychological - personality (see Figure 2). However, acknowledging that scientists may not agree on the nesting of variables and future research may evolve the interplay, the taxonomy is flexible to accommodate reclassifications without re-coding. The current taxonomy includes approximately 5,000 entries (a highly abbreviated version is depicted in Figures 2, 3, and 4).

The process of taxonomy development began by extracting correlation matrices from of two top-tier journals in organizational research (Journal of Applied Psychology and Personnel Psychology) from 1980-2010. From these extracted data, subject matter experts classified the variables in an unconstrained manner. Importantly, no single taxonomic classification scheme will match mental models held by every organizational researcher because a great deal of variance likely exists in mental models across researchers. Thus, we are currently pursuing two approaches to validation. First, we will apply a series of agreement assessment approaches (e.g., Q-sort, constrained sorting) to allow the specification of an ideal-fitting taxonomy. Second, we will apply advanced statistical approaches such as cluster analysis, using the taxonomy and accumulated research findings, to determine the best data-driven taxonomic arrangement.

D. Technology

During our preliminary efforts, we used a combination of PDF extraction software (Able2Extract; commercial), mind-

mapping software (Freemind; open-source), and Microsoft Excel to manage the database. We have planned in advance for expansion of our project by choosing mind-mapping software that is able to export hierarchies in an easily-transportable format (e.g., XML). In addition, we have adhered to basic database management practices in our Excel database, including unique identifiers for data rows and unique delimiters for isolating and migrating data. As the project moves forward, we plan to create advanced coding forms, database management tools, and GUIs using a combination of open-source database software such as MySQL and the Java Virtual Machine (JVM) using a modern software stack as that managed by http://www.typesafe.com. Further, technologies enable the seamless extraction of text from PDFs in online systems and the translation of PDFs to HTML5 documents using libraries such as http://www.idrsolutions.com/java-pdf-library/ are now available.

The coding software automatically extracts basic article data (e.g., title, authors, and year) from an academic bibliographic database. Built in error checks minimize coding problems and variables are selected from pull-down menus. Psychometric corrections are programmed into the software, allowing easy implementation. The framework can be updated and customized for specific meta-analytic projects, with the data saved as a flat file. More advanced technological developments for coding, classification, and database management are currently underway.

III. Discussion

As the project moves forward, we will integrate the taxonomy with the coding software, so that coders can choose subsets of constructs to code, including by custom or default lists and by keyword search. This will require using a relational database structure. We seek to migrate and enhance the implementation of the features already included in our legacy software, such as psychometric corrections for unreliability. The resulting database will be able to identify who coded which articles, whether they have been coded by multiple coders, or only partially coded.

As our goals are realized, we will be increasingly able to address a host of long-standing and central research questions in our field. With more than one million findings linked to a hierarchical taxonomy and meta-analytic software, we will be able to ask questions such as: (a) Overall, how well do psychological states and traits predict employee behaviors such as performance and turnover?, (b) Across a variety of domains, does the intention-mediated model (i.e., attitude - intention - behavior; [32]) actually fit our cumulative observations?, (c) Have our effect sizes declined over time?, (d) Do sample types and response rates matter (i.e., influence effect sizes) for different major types of relations?, (e) What are the actual, empirical effect size benchmarks across a variety of relation types (e.g., What elements of a performance management program combine to create the greatest employee performance and engagement? How effective does a training program need to be in order to be above average?), (f) What is the actual, observed relation between reliability and validity across a variety of topics; are effect sizes over-corrected in meta-analysis?, (g) What selection tools work best in different industries located in different parts of the world?, and (h) To what extent does empirical redundancy appear in the HRM literature (e.g., are a variety of supervisor attitudes actually empirically distinguishable, in terms of their interrelations and relations to other constructs)? Clearly, the set of possible research questions to address both researcher and practical interests is vast. However, these pressing questions cannot be answered and continuously updated in an efficient manner with current meta-analytic approaches.

IV. Conclusion

We have described an ongoing Digging into Data Challenge project called metaBUS. Our near-term goal is for the metaBUS Project to change the speed of science and the nature of collaboration in the field of human resources and organizational behavior with the eventual goal of extending the mapping and corpus of data to far-reaching fields.

References

[1] F. L. Schmidt, "What do data really mean? research findings, meta-analysis, and cumulative knowledge in psychology," *American Psychologist*, vol. 47, no. 10, pp. 1173–1181, 1992.

[2] S. Kepes, G. C. Banks, M. McDaniel, and D. L. Whetzel, "Publication bias in the organizational sciences," *Organizational Research Methods*, vol. 15, no. 4, pp. 624–662, 2012.

[3] H. Aguinis, D. R. Dalton, F. A. Bosco, C. A. Pierce, and C. M. Dalton, "Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact," *Journal of Management*, vol. 37, pp. 5–38, 2010.

[4] J. A. Colquitt, D. E. Conlon, M. J. Wesson, C. O. Porter, and K. Y. Ng, "Justice at the millennium: A meta-analytic review of 25 years of organizational justice research.," *Journal of applied psychology*, vol. 86, no. 3, pp. 425–445, 2001.

[5] J. A. Colquitt, B. A. Scott, J. B. Rodell, D. M. Long, C. P. Zapata, D. E. Conlon, and M. J. Wesson, "Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives.," *Journal of Applied Psychology*, vol. 98, no. 2, pp. 199–236, 2013.

[6] J. A. Shaffer and B. E. Postlethwaite, "A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures," *Personnel Psychology*, vol. 65, no. 3, pp. 445–494, 2012.

[7] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: A meta-analysis," *Personnel psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[8] J. Gleick, *The Information: A History, a Theory, a Flood*. Fourt, 2011.

[9] T. A. Judge, C. J. Thoresen, J. E. Bono, and G. K. Patton, "The job satisfaction-job performance relationship: A qualitative and quantitative review.," *Psychological bulletin*, vol. 127, no. 3, pp. 376–407, 2001.

[10] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Communications of the ACM*, vol. 30, no. 11, pp. 964–971, 1987.

[11] H. Le, F. L. Schmidt, J. K. Harter, and K. J. Lauver, "The problem of empirical redundancy of constructs in organizational research: An empirical investigation," *Organizational Behavior and Human Decision Processes*, vol. 112, no. 2, pp. 112–125, 2010.

[12] K. Leavitt, T. R. Mitchell, and J. Peterson, "Theory pruning: Strategies to reduce our dense theoretical landscape," *Organizational Research Methods*, vol. 13, no. 4, pp. 644–667, 2010.

[13] S. L. Rynes, K. G. Brown, and A. E. Colbert, "Seven common misconceptions about human resource practices: Research findings versus practitioner beliefs," *The Academy of Management Executive*, vol. 16, no. 3, pp. 92–103, 2002.

[14] F. L. Schmidt and J. E. Hunter, "The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings.," *Psychological bulletin*, vol. 124, no. 2, pp. 262–274, 1998.

[15] K. Börner, R. Klavans, M. Patek, A. M. Zoss, J. R. Biberstine, R. P. Light, V. Larivière, and K. W. Boyack, "Design and update of a classification system: The ucsd map of science," *PloS one*, vol. 7, no. 7, p. e39464, 2012.

[16] K. Börner, C. Chen, and K. W. Boyack, "Visualizing knowledge domains," *Annual review of information science and technology*, vol. 37, no. 1, pp. 179–255, 2003.

[17] J. D. Novak and A. J. Canas, "The theory underlying concept maps and how to construct and use them," *Florida Institute for Human and Machine Cognition*, vol. 284, 2008.

[18] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 1, pp. 17–30, 1989.

[19] K. W. Boyack and R. Klavans, "Creation of a highly detailed, dynamic, global model and map of science," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 670–685, 2014.

[20] I. Zupic and T. Cater, "Bibliometric methods in management and organization: A review," tech. rep., Paper presented at the annual meeting of the Academy of Management, Orlando, FL, 2013.

[21] R. P. Leone, L. M. Robinson, J. Bragge, and O. Somervuori, "A citation and profiling analysis of pricing research from 1980 to 2010," *Journal of Business Research*, vol. 65, no. 7, pp. 1010–1024, 2012.

[22] Z. Ma, D. Liang, K.-H. Yu, and Y. Lee, "Most cited business ethics publications: mapping the intellectual structure of business ethics studies in 2001–2008," *Business Ethics: A European Review*, vol. 21, no. 3, pp. 286–297, 2012.

[23] M. Shafique, "Thinking inside the box? intellectual structure of the knowledge base of innovation research (1988–2008)," *Strategic Management Journal*, vol. 34, no. 1, pp. 62–93, 2013.

[24] C. I. Lee, W. Felps, and Y. Baruch, "Toward a taxonomy of career studies through bibliometric visualization," *Journal of Vocational Behavior*, in press.

[25] J. Higgins, "Green s. cochrane handbook for systematic reviews of interventions. version 5.1. 0. the cochrane collaboration; 2011," 2013.

[26] A. Shlonsky, E. Noonan, J. H. Littell, and P. Montgomery, "The role of systematic reviews and the campbell collaboration in the realization of evidence-informed practice," *Clinical Social Work Journal*, vol. 39, no. 4, pp. 362–368, 2011.

[27] S. M. Wolf, B. N. Crock, B. Van Ness, F. Lawrenz, J. P. Kahn, L. M. Beskow, M. K. Cho, M. F. Christman, R. C. Green, R. Hall, *et al.*, "Managing incidental findings and research results in genomic research involving biobanks and archived data sets," *Genetics in Medicine*, vol. 14, no. 4, pp. 361–384, 2012.

[28] R. Veenhoven, "World database of happiness. example of a focused findings archive." RatSWD Working Paper, 2011.

[29] S. Ip, N. Hadar, S. Keefe, C. Parkin, R. Iovin, E. M. Balk, and J. Lau, "A web-based archive of systematic review data," *Syst Rev*, vol. 1, no. 1, pp. 15–22, 2012.

[30] A. S. Levi, "Humanities big data: Myths, challenges, and lessons," in *IEEE International Conference on Big Data*, pp. 33–36, October 2013.

[31] W. F. Cascio and H. Aguinis, "Research in industrial and organizational psychology from 1963 to 2007: changes, choices, and trends.," *Journal of Applied Psychology*, vol. 93, no. 5, pp. 1062–1081, 2008.

[32] M. Fishbein and I. Ajzen, *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley, 1975.