Metadata Analytics: A Methodological Discussion

Jian Qin[1] Sarah Bratt[2] Jeff Hemsey[1] and Alexander O. Smith[1]

[1] *{jqin, jjhemsle, aosmith@syr.edu}*

School of Information Studies, Syracuse University, Syracuse, NY 13244 (USA)

[2] *sebratt@arizona.edu*

School of Information, University of Arizona, Tucson, AZ 85721 (USA)

Abstract

Metadata Analytics is a term used to describe a research field that utilizes quantitative methods and metadata for publications, patents, datasets, and other research entities to study science of science. Metadata analytics inherits the bibliometric and scientometric tradition while infusing novel data sources – metadata for datasets – to extend the traditional bibliometric and scientometric research. The large scale of metadata from scientific data repositories offers both opportunities and challenges in the quantitative study of science. This paper discusses the problems and opportunities that metadata analytics contends with from a methodological perspective. Using the authors' experiences over the course of a multi-year metadata analytics project, the paper focuses on the subtle differences between methods and science (or means and end) that arise when conducting research in metadata analytics and, for the same reason, bibliometrics and scientometrics . Metadata analytics is both a methodology and a research field. The intertwining of methods and science in metadata analytics can create pitfalls for researchers. Steering clearly between the means and ends in metadata analytics is essential to produce good science.

Introduction

Research in the science of science can be generalized in three broad areas: the science of career, the science of collaboration, and the science of impact (Wang & Barabási, 2021). The publication metadata – articles and their authors, subjects, and citations – have been the primary sources for establishing metrics and models for exploring the patterns, evaluating the impact, and predicting the trends of the scientific research enterprise. Using publication metadata exclusively for studying the science of science reflects the academic evaluation culture that gives a heavier weight to publications than to other research activities such as data collection and processing. The fact that scientific journal articles represent the end-product of a research lifecycle leaves a data gap due to a reliance on publication-centric metadata. As a result, we exclude most research activity taking place prior to publications. This data gap is problematic because it neglects an important asset in data-intensive science: datasets and the authors related to the creation of these datasets. In the age of "fourth paradigm" science, scientific data production is increasingly recognized as part of scholarly output, as exemplified by data journals such as Nature's *Scientific Data* and the *Biodiversity Data Journal*. Clearly, without capturing the full scientific lifecycle – including data production and its authorship – the study of the science of science would be incomplete and vulnerable to bias.

In the last decade, the Metadata Lab team at the School of Information Studies, Syracuse University, has been trying to address this data gap in the science of science research. We have been using the metadata in GenBank records to study the structures and dynamics of collaboration networks of the global research community and developed a theoretical framework of Collaboration Capacity (CC) (Qin et al., 2018; Hemsley et al., 2022). The CC framework incorporates data submission metadata together with the associated publication metadata that span from the debut of GenBank in 1984 until the present time (our last GenBank metadata update contains data up to 2021) and established new metrics to reflect the dynamics of the data to knowledge process (Hemsley et al., 2020). We use a term *Metadata Analytics* to represent this new approach that utilizes metadata from scientific data and/or publication

repositories to study the career of scientists, collaboration networks, knowledge diffusion patterns, and the state of the scientific research enterprise.

The notion of *Metadata Analytics* stands both as a research field of its own right and as a methodology. As a research field, its history can be traced back to the 1960's when the exponential growth of scientific literature triggered the advent of *Science Citation Index* and a wide range of research topics under the names of bibliometrics, scientometrics, and informetrics. The measures and metrics used in these analyses have produced bibliometric "laws" and statistical models such as the Bradford Law of Scattering of scientific literature (Summers, 1983) and the Lotka's law of scientific productivity (Lotka, 1926; Bookstein, 1977). The availability of big data (as well as big metadata) and rapid advances in computational methods afford metadata analytics opportunities to move from empirical, explanatory to be more exploratory and theoretical. As a methodology, metadata analytics employs quantitative approaches to study the phenomena of science workforce and careers, scientific productivity, collaboration networks, data and knowledge transfer, and innovations. Simple frequency counts of authors, papers, and citations have been transformed into sophisticated measures to address the multi-facets and complexity in evaluating the science research enterprise. Adding metadata from data repositories on top of these changes, we see a need for a more inclusive label for the quantitative study of data-intensive science today.

While metadata from large data repositories have been proved to be able to offer new insights into the collaboration networks and advance our understanding of the science of science, there is a heavy cost in making these metadata as well as other associated data sources readily analyzable due to the data formats that are not easily interoperable nor friendly for scientometric analysis. At a macrolevel, the large volume of repository metadata creates a burden for data parsing, cleaning, and transformation. If additional data are collected and merged, e.g., funding data from NIH RePORTER, there will be an extra, non-trivial burden and cost in data wrangling and merging to connect metadata from different repositories. The need to constantly update the metadata collected causes repetition in data processing because most computational codes are not reusable without substantial maneuvering.

This paper discusses metadata analytics from a methodological perspective. The goal is to address the data problems and review the approaches used in metadata analytics. Discussions of this type are important because the boundaries between metadata analytics as a methodology and as a field of research are not always clear-cut and can cause confusions between the means and ends of metadata analytics. By laying out the similarities and differences between the two aspects, we hope to establish a clearer understanding of metadata analytics and avoid pitfalls and mistakes that could impact the research reliability and validity. This paper is not intended to conduct a comprehensive review of all methods used in bibliometrics and scientometrics, rather, it attempts to clarify the subtle differences between the dual aspects of metadata analytics—as a field of its own right and as a method. The rest of this paper will be organized in the following sections: review of metadata analytics methods, big metadata in data repositories, experience from GenBank metadata analytics project, and conclusion.

Review of Metadata Analytics Methods

Bibliometrics

Using metadata to analyze scientific research enterprise is not new and has been around for as long as the exponential growth of scholarly literature started in the early 1960's. A better-known term is *Bibliometrics*, which is generally considered as a kind of method and measurement for studying literature and its authors. This term was first used by Alan Pritchard who defined bibliometrics as the "application of mathematics and statistical methods to books and other

media of communication" (Pritchard, 1969, p. 349). Bibliometrics was also called the "scientific study of recorded discourse" (Schrader, 1981). Bibliometric analyses can take a relational or evaluative approach. The former answers research questions such as "Who is related to whom?" Relational bibliometric analyses tend to be descriptive and explanatory and can be applied to study individuals, organizations, domains, or geographical areas. The latter – evaluative bibliometric analyses – is commonly used to assess "the level of quality, importance, influence, or performance of individual documents, people, journals, groups, domains (subject areas, fields, or disciplines), or nations" (Borgman and Furner, 2002, p. 11).

Measures used in both explanatory and evaluative studies may vary widely, but in essence they are derived from frequency counts of major research output types and entities associated. We use "unit of analysis" and their applications in four research areas of science of science in Table 1 to construct a basic understanding of how metadata units are used. These units of analyses can be coordinated, combined, and transformed to generate or formulate many sophisticated measures according to research questions and purposes. Historically, bibliometrics has used the publication metadata exclusively as the unit of analysis. Whether it is the h-index (Hirsch, 2005), Q-Model (Wang & Barabási, 2021), or invisible college (Crane, 1972), these metrics are derived from the products of a "Publish or Perish" academic culture. The role and contribution of metadata for datasets in scientific research lifecycle was either taken for granted or not in the equation of bibliometrics.

Table 1. Unit of analysis vs. Applications in metadata analytics

Unit of analysis	*Applications*			
	Career	*Collaboration/ Teams*	*Knowledge/ Innovation*	*Impact*
Publication count	X	X		
Citation count to publications			X	X
Publication authors	X	X		
Citation count to datasets			X	X
Data authors	X	X		

Citation-based measures, for example, are often used to explain and evaluate citing and cited relations among publications and rank authors, papers, and journals based on their positions in citation networks. The measures shown in Table 2 summarize the 39 measures listed in Bollen et al. (2009) from a methodological perspective. These measures are derived from citation counts that have been processed and transformed into statistical probabilities (e.g., journal use and cite probabilities) and network properties (e.g., degree and betweenness centralities). Although Bollen et al. (2009) call the 39 measures as "scientific impact measures," they have inherent limitations in depicting the whole picture of impact, which have been discussed extensively in past research (MacRoberts & MacRoberts, 1989).

Table 2. Existing measures for scholarly impact (compiled according to Bollen et al., 2009)

Function of measures	Type	
	Citation	Usage
Ranking	Scimago Journal Rank, PageRank, Y-factor	PageRank
Citedness	Cites per doc, Journal Impact Factor, Scimago Total Cites, Journal Cite Probability	Journal Use Probability, Usage Impact Factor

Relation	Closeness centrality, out-degree centrality, degree centrality, in-degree centrality, betweenness centrality	Closeness centrality, degree centrality, in-degree centrality, betweenness centrality, out-degree centrality
Index	Immediacy index, H-index, citation half-life	

(Source of Table 2: Qin, 2010)

Another important area of bibliometric study is the evaluation of people (Borgman & Furner, 2002). Number of papers, venues of published papers, and citation counts are typically used to assess a faculty member's productivity, quality, and influence of her/his research outputs in the tenure and promotion process, despite the limitations identified in literature (MacRobers & MacRoberts, 1989). Co-author counts are the foundation to construct collaboration maps to explain researchers' positions in collaboration networks and the nature of their connections with collaborators. The study of co-authorship networks combines the publication co-author and/or citation/co-citation counts and network science methods to discover and model the network shapes and structures as well as scholarly communities (Kumar, 2015).

The increasing attention to data-intensive science brings new developments in bibliometric studies. Citations to datasets have been used to measure the use and infer the impact of datasets (Silvello, 2018). Publication citation networks have been used widely in scientometric analyses because they indicate "crediting an idea, signaling knowledge of the literature, or critiquing others' work" (Martyn, 1975), but datasets citation is advancing, however slowly (e.g., Zeng, et al., 2020). A few studies have begun to incorporate metadata from data submissions and associated publications to measure the impact of scholarly products other than publications (e.g., Belter, 2014), study scientific collaboration networks (e.g., Li, et al, 2022), and estimate data use statistics (e.g., Robinson-Garcia, et al., 2017). The expansion of data sources for bibliometric studies has gone beyond the publication metadata to include metadata for datasets, funded grants, and patents. As data-intensive science generated more data than ever before and metadata for datasets in open repositories are being used in metadata analytics, questions arise about the new label *Metadata Analytics*: Is metadata analytics a methodology or a science? Is there any difference between metadata analytics and bibliometrics? How will the inclusion of metadata for datasets affect the quantitative study of science?

The term *metadata analytics* emerged in this context to represent a research area that applies computational and network science methods and techniques to the analysis of metadata from scientific data repositories (Qin et al., 2018). The fact that publication metadata is no longer the only data source for quantitative study of science warrants the use of metadata analytics to be more inclusive of new data sources. As such this term carries more weight on the side of methodology than on the side of a science because even though the empirical findings from metadata analytics may be able to derive statistical or mathematical models and indices, the explanation and interpretation of empirical findings and statistical models often need the help of social sciences theories and methods to make sense of the findings and models.

Network Science

Network science is a popular approach used in bibliometric analysis to investigate networks of authors and citations for a variety of purposes, ranging from collaboration patterns and trends and community detection to idea development paths to innovation tracking. Research collaboration is typically measured by coauthorship in publications and at the international, interinstitutional, interdepartmental, or departmental level. Researchers in a collaboration network are called nodes or vertices and the relationships (i.e., coauthorship) between nodes are edges. Collaboration networks with very large numbers of nodes and edges together with variant weights of edges and other factors are highly complex as nodes have uneven numbers

of edges and the edges may vary in length between nodes. Such networks consist of clusters or communities of researchers, which are self-organized, may be interconnected in some ways, and evolve over time (Newman, et al., 2006).

Barabási et al. (2002) provides a summary of the characteristics of collaboration networks, which include: 1) most networks have the "small world" property, 2) real networks have an inherent tendency to cluster, more so than comparable random networks, and 3) the distribution of the number of edges for nodes (degree distribution) "contains important information about the nature of the network, for many large networks following a scale-free power-law distribution" (p. 591). These network theories and models have been applied in studying collaboration networks in biology, ecology, and physics. Several properties of scientific collaboration networks have been identified in these studies: small worlds are common in scientific communities; the networks are highly clustered; and biomedical research appears to have a much lower degree of clustering compared to other disciplines such as physics (Newman, 2001). Studies of the evolution of scientific collaboration networks shows that the degree distribution follows a power law and key network properties (diameter, clustering coefficient, and average degree of the nodes) are time dependent, that is, the average separation decreases in time and clustering coefficient decays with time (Barabási et al., 2002).

Network science as a method has been applied in citation analysis as well. The metrics listed in the Relation category in Table 2 are typical statistical properties in measuring positions, change patterns, and connectedness of nodes in networks. For example, citation network clusters can reveal how research specialties transformed and changed into stand-alone fields over time (Rosvall & Bergstrom, 2010). Citation or co-citation networks are the major approaches in detecting and visualizing patterns and trends in research fields, examples of which include the scientometric analysis and visualization of research activities in the architecture, engineering, and construction industry (Darko et al., 2020) and a co-citation analysis of emerging trends and new developments in information science (Hou et al., 2018). The versatility of network science makes it a popular approach in bibliometric and scientometric analyses for studying not only collaborative relations among authors but also the "formal and informal flows of information, ideas, research practices, tools, and samples" (Fortunato et al., 2018, p. 3). Network science is a research field of its own right yet can be applied in almost all aspects of science of science research.

Big Metadata in Research Data Repositories

The emergence and evolution of data-intensive science not only changed the way science is conducted but also created a vast amount of new data sources that can be utilized to study the science enterprise. Scientific data have grown at an exponential rate just as what scientific literature's growth experienced after the World War II (Meadows, 1998; de Solla Price, 1986). The data repositories together with software applications form an important part of the cyberinfrastructure (CI) serving the data-intensive science research. There are abundant examples demonstrating the large scale research networks enabled by the CI environment: the use of CI allows researchers to integrate data from multiple observatories in the Laser Interferometer Gravitational Wave Observatory (LIGO) experiments, to gather and analyze data over spatial and temporal dimensions in the Long Term Ecological Research Network (LTER), and share and update genetic sequencing data in the National Center for Biotechnology Information (NCBI) data repository GenBank. In all three examples mentioned here, research collaboration enabled by CI have expanded to an unprecedented scale, either directly as reflected in long lists of authors in publications or indirectly as reflected in the fast-growing sizes of data and publication repositories. These advances in science have outpaced our ability to develop quantitative models and metrics to analyze CI-enabled science enterprise. In this

paper we use GenBank and its related data repositories at NCBI to illustrate the value of metadata in these data repositories and why they can and should be included in modern metadata analytics for the study of the science of science.

NCBI Data Repositories

NCBI data repositories curate data on genomic projects, biosamples, molecular sequences, chemicals, and bioassays, as well as the software tools for finding, identifying, selecting, obtaining, and exploring these biomedical data (Sayers et al., 2021). GenBank is part of the International Nucleotide Sequence Database Collaboration (https://www.insdc.org/) that curates publicly available genetic sequences with annotations. As the largest nucleotide data repository in the world, it contains sequences from all branches of life and is considered a foundation for medical and biological discovery (Bloom et al., 2021).

The role of GenBank in medical and biological discovery makes it critical to link the sequence data to other relevant repositories. Many annotation records in GenBank have been linked to other related data archives or repositories. For example, a GenBank record may contain IDs from BioProject and BioSample databases to allow for tracking from which BioProject and BioSample the genetic sequences were generated. A record in ClinVar (a database for genomic variations that affect human health) contains GenBank accession number(s) that allow researchers to traverse multiple databases to gather related genomic data and track developments.

Sequence data submitted to GenBank provide information on the sequence as well as the references (i.e., publications) associated with the sequence data (see Figure 1 for a sample record). The direct links between authors of sequence data and the publications add not only a new data source beyond the publication metadata but also opportunities to re-examine the conventional measures in which the data production metadata have been absent. As scientific research becomes increasingly data-intensive, there is every reason for us to consider the data author and submissions when studying the science of career, collaboration, and impact.

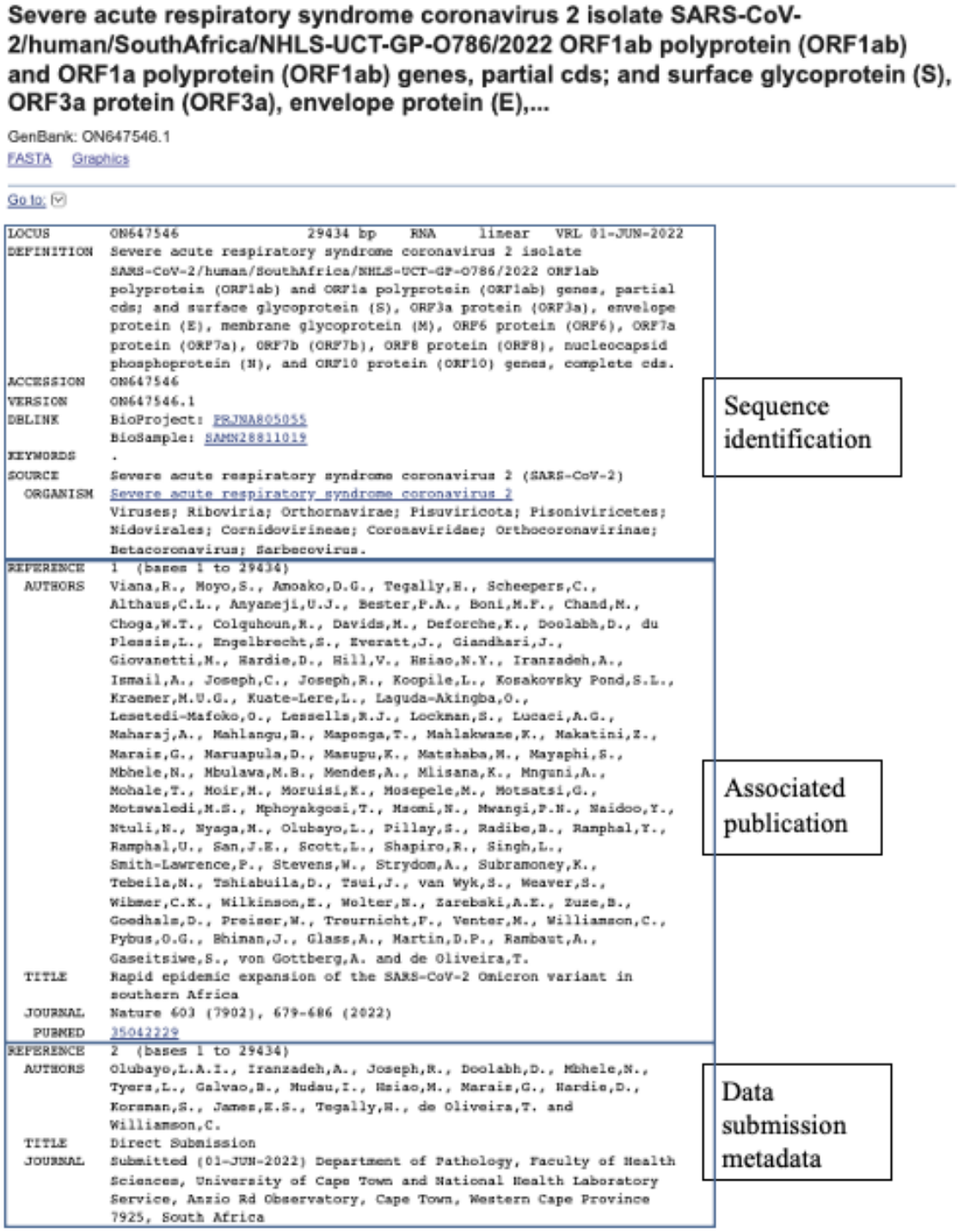

Figure 1. The metadata section in a GenBank annotation record

(source: https://www.ncbi.nlm.nih.gov/nuccore/ON647546)

Other Data Sources

In addition to metadata from open research data repositories, there are a wide variety of other data sources that might be able to be used for triangulation or primary analysis. For example, the NIH RePORTER, Patents, NSF statistics, MS Academic Graph, Semantic Scholar, and Commercial databases such as Web of Science, Elsevier Scopus, and IRIS UMetrics, and Open Alec are rich data sources that can offer contextual information about datasets.

Challenges in Using Repository Metadata as a Data Source

Just as research data are pivotal to modern science, metadata describing research data are also pivotal for data discovery, sharing, reuse, crediting authors, and reproducing the research. Longitudinal metadata can also offer insights into the history of science advances and the development of science enterprise at national, disciplinary, and institutional levels. The metadata for data authors and datasets only recently started to be included in the study of science of science (Costa et al., 2016; Qin, Hemsley, & Bratt, 2022). While opportunities in utilizing metadata from data repositories as a research data source are exciting and promising, there are major obstacles in using them to conduct research because of their inherent complexity (Bratt et al, 2017), messy, ambiguous, and unstructured nature (Chen & Sarkar, 2011). Idiosyncratic structures and formats in data repositories vary greatly from discipline to discipline, which make data fusion and integration extremely challenging for non-data science researchers. A significant investment of skilled professionals, computing resources, and funding is needed for fully realizing the value of metadata to enable data-intensive interdisciplinary research.

Aside from the problems in data structures and readiness for analysis, metadata from data repositories can often blur the boundaries between science and methodology, e.g., network science is a method for collaboration network analysis but at the same time it is also a research field itself. The blurring boundaries between the study of science and the methodologies for studying science can confuse the end and means in metadata analytics.

Metadata analytics often needs to collect metadata from multiple types of repositories in addition to those for curating scientific data. Metadata about patents and funding, for example, can be combined with metadata from GenBank to study scientists' career trajectories and collaboration networks as well as the impact of enablers for collaboration capacity (Hemsley, Qin, & Bratt, 2022). The triangulation of these data can offer more insightful analysis on collaboration networks, research performance, funding, and knowledge diffusion.

Experience from GenBank Metadata Analytics Project

The GenBank metadata analytics project started in 2012 with a pilot test for exploring the feasibility of including data submissions in the science of science research. This project started with a goal to investigate the research network structures and dynamics emerging around cyberinfrastructure. The metadata we collected from GenBank initially span from its inception in 1984 to 2013 and later had two updates that cover up to May 2021.

The metadata section in a GenBank annotation record has the function of identifying the genetic sequence, linking associated references with the sequence, and documenting the submission information (Figure 1). NCBI releases GenBank data in compressed files via an FTP server on a quarterly basis. The most recent GenBank flat file release (251, released on August 15, 2022) consisted of 5,836 compressed files in a total 678,133 MB with a compress ratio approximately 20%. The number of files and size of data volume require the whole workflow from file downloading to extraction to wrangling to be performed by computer programs.

A challenge that has plagued much of bibliometric and scientometric analysis is the disambiguation of author names. The GenBank metadata records were no exception. The GenBank metadata contains three categories of author names: publication author, dataset contributor, and patent inventor. Fortunately for data cleaning, the names are represented in a standard format: Surname, First Initial, Middle Initial (e.g., Börner, K.). However, no unique author identifiers (e.g., ORCIDs) exist to disambiguate the authors. To address this issue, we used the 2013 KDD Cup Data Mining Contest solution (Liu et al., 2013) and SCOPUS author data to disambiguate the publication and dataset scientists' names (96% accuracy). The patent inventor names were disambiguated with the KDD solution plus the U.S. Patent and Trade-

mark Office (USPTO) database author names (97% accuracy). The author name disambiguation challenge will persist if author names are not uniquely identified. The author name disambiguation task is a temporary solution to a broader problem. For instance, the disambiguation task must be completed each time when we update the GenBank metadata records. New author names are added to the collection of metadata, each of which requires disambiguation to ensure there is not an overcounting or undercounting of the authors.

Even though affiliation and other author metadata can be leveraged in author name disambiguation, they are not readily available to be used by the disambiguation algorithm. The author affiliation and geographic information (Figure 1) in a GenBank dataset submission is in the JOURNAL field and includes the name of the author's university/company, and often the department or laboratory associated with the scientist. The geographic metadata includes the country in which the scientists' university is housed and often the mailing address and/or postal code. When multiple authors from different institutions are credited in a data submission, the current GenBank data structure does not provide clear links between authors and their affiliations. We had to use the PubMed ID in the GenBank record, where author names are linked to their institutions, to match authors with their institutions in data submission metadata when the need arose.

As an early explorer of using metadata from a scientific data repository for scientometric analysis, we gained a better understanding of this novel data source for the science of science research and learned two main lessons from working with this very large scale of metadata.

Methods vs. Science (Means vs. Ends)

The science of science research is quantitative by nature with the goal to understand the interactions among scientific agents and conditions underlying creativity and scientific discovery (Fortunato et al., 2018). In the process of cleaning and transforming data and creating datasets for addressing our research questions, we heavily used statistical and network science methods. One question we faced most frequently was: What is exactly we are doing: developing bibliometric/scientometric/network science theories or doing science? By doing science, we mean applying some methods to analyze the data, interpret the results based on some theories, and derive some conclusions (and possibly build new theories). For example, the analysis results show that GenBank publication author networks follow a power law distribution and possess the characteristics of scale free networks. This finding presents a trajectory that can leads our discovery in two directions: claim a success in proving the GenBank's scale-free network characteristics or explore applicable theories for interpreting this phenomenon. Discussions at our team meetings clarified that network science is used in this metadata analytics project as a method. In other words, the network science theories and techniques are the means for our project; the end of our project is to conduct science of science research to uncover collaboration dynamics, patterns, and interactions. The fact that these empirical findings supported models and theories in network science opens the door for interpreting the findings using richer theories from social studies of science.

The lesson learned from clarifying the boundaries between methods and science is that metadata analytics is a means through which we pursue the goal of studying the science of science. However, the boundaries are not always clear-cut. A decline in clustering coefficient over time may implicate a novel property of scale free networks in the context of our data, that is, the whole network evolved over time from a small number of "hubs" into a more distributed, flatter network with more smaller clusters. While this explanation may be sufficient for network science, it is not for our project because this network science explanation cannot answer the question of "so what?" In other words, the mathematical language is not the goal of our research, rather, it is a tool for diagnosing the data to allow quick identification of patterns and

trends, but what these patterns and trends mean for the science research enterprise would require deeper exploration so that we can confidently answer the "so what" questions.

Theoretical Development

The massive amount of GenBank metadata for both data submissions and publications created an unprecedented opportunity to quantitatively study the bench (or basic) biomedical research enterprise (vs. bedside or clinical research). While the data analysis generated statistical properties and visualizations of the networks, it is important to keep in mind that all these are empirical phenomena. How to capture and elaborate the epistemological implications from these empirical phenomena becomes the first step toward theory development. We took two approaches in this process: building a conceptual framework by incorporating related theories and developing new metrics based on the new data sources.

Based on our empirical findings and inspired by Bozeman, Dietz, and Gaughan (2001)'s Scientific and Technical (S&T) Human Capital model, we used a term "collaboration capacity" to refer to the ability of an individual researcher or a team of researchers to collaborate throughout the data production and publication lifecycle and sustain a network of collaborators over time. We built an initial "collaboration capacity framework" in Qin, Hemsley, and Bratt (2018), which was based on three assumptions: 1) collaboration capacity is a proxy for studying scientific capacity, 2) multi-stages (data, publication, and patent) of a research lifecycle can be used as a proxy for studying knowledge diffusion, and 3) collaboration capacity has impact on the level of research productivity and extent of knowledge diffusion (Figure 1).

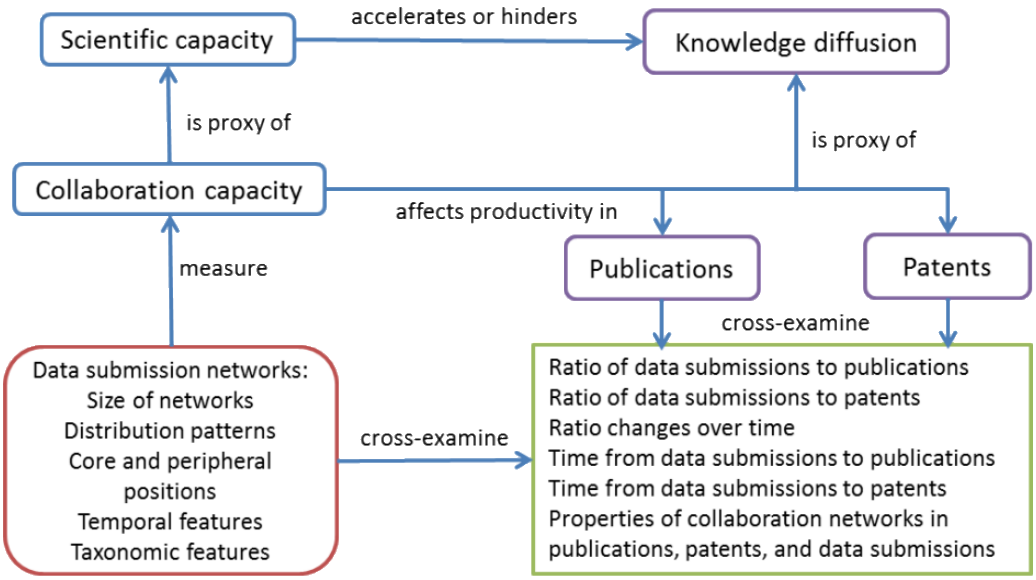

Figure 1. The collaboration capacity framework version 1.0 (Qin, Hemsley, & Bratt, 2018)

As the data analysis unfolded and the funding data from NIH incorporated, the initial framework appeared to be insufficiently narrow. On a theoretical perspective, collaboration capacity is enabled primarily by three things: the S&T human capital, cyberinfrastructure, and science policy. The S&T human capital is the sum of scientific, technical, and social knowledge, skills and resources embodied in a particular individual (Bozeman, Dietz, & Gaughan, 2001). Cyberinfrastructure includes data and publication repositories, software tools, and discover services supporting scientific research. Science policy ensures resource allocation and dissemination of research outputs among other things. They are the three "enablers" of collaboration capacity. Whether a scientist has a strong or weak collaboration capacity, therefore, is dependent on how effective the enablers are in helping strengthen the collaboration capacity of an individual, team, or institution. On an operationalization level, the measurement

of one's collaboration capacity can be grouped into two categories: data production and data-to-knowledge metrics (Figure 2).

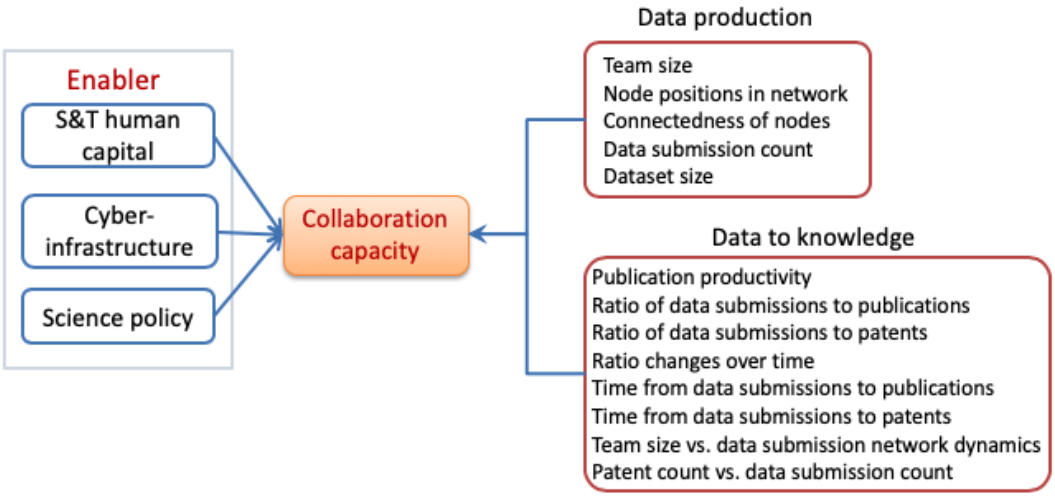

Figure 2. The collaboration capacity framework version 2.0

The right side of Figure 2 are measures derived from both publication and dataset metadata, representing the quantitative, empirical aspect of collaboration capacity, while the left side is more theoretical. Although we have developed some metrics to operationalize the impact and effectiveness assessment of CC enablers, they are still under testing and evaluation. Other theories such as the knowledge diffusion and innovation (Morone and Taylor, 2010) may be useful for modeling and formulating the components in this model and interpreting the findings. There is still much to be explored for the version 2.0 of this framework. For example, there are close relations among the three enablers: allocation of funding for training can grow S&T human capital, which can in turn strengthen the cyberinfrastructure's capacity in supporting scientific research. Collaboration capacity mirrors the effectiveness of enablers. In this sense, measuring collaboration capacity provides primary data to measure the effectiveness and impact of enablers of collaboration capacity.

Conclusion

In this paper we discussed the connotation of metadata analytics and its relations to bibliometrics and scientometrics. The use of metadata from data repositories promises new perspectives and opportunities in developing new theories and metrics for the science of science. Yet, the inherent problems related to data formats induce high costs in data collection, cleaning, processing, and aggregation with other data sources. Some of these problems can be resolved by disruptive measures, e.g., assigning standard identifiers to authors forwardly and retrospectively for both publications and datasets, which many journal publications already started implementing. Others such as abbreviated author name format in database records would require changes in conventions and practices from the top down. Standardization in data formats and author identification will be the foundation for building a data infrastructure for science of science research. Relabeling this research field with "metadata analytics" is to raise awareness of and develop appropriate strategies to address these issues in the science of science research community so that a data ecosystem can be built to support the research and community.

Metadata analytics is both a methodology and a research field. It inherits the methods and practices that bibliometrics and scientometrics have produced but uses a wider variety of data sources than bibliometrics and scientometrics have traditionally used. The fact that metadata analytics encompasses metadata from data repositories creates a possibility to quantitatively study the science enterprise starting from data production (data submissions) to knowledge

creation (publications) to innovations (patents). The intertwining of methods and science in metadata analytics can have hidden pitfalls for researchers. Steering clearly between the means and ends in metadata analytics is essential to produce good science.

Acknowledgement

Research reported in this publication was supported by National Science Foundation Award No. 1561348 and the National Institute of General Medical Sciences of the National Institutes of Health under Award No. R01GM137409. The content is solely the responsibility of the authors and does not necessarily represent the official views of National Science Foundation and the National Institutes of Health.

References

Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaboration. *Physica A*, 311: 590-614.

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS One*, 9(3), e92590. https://doi.org/10.1371/journal.pone.0092590

Bloom, D. E., Cadarette, D., Ferranna, M., Hyer, R. N., & Tortorice, D. L. (2021). How new models of vaccine development for covid-19 have helped address an epic public health crisis. *Health Affairs*, 40(3), 410-418. https://doi.org/10.1377/hlthaff.2020.02012

Bookstein, A. (1977). Patterns of scientific productivity and social change: A discussion of Lotka's Law and bibliometric symmetry. *Journal of the American Society for Information Science*, 28(4): 206-210.

Borgman, C.L. & Furner, J. (2002), Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, 36: 2-72. https://doi.org/10.1002/aris.1440360102

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. PLoS ONE 4(6): e6022. https://doi.org/10.1371/journal.pone.0006022

Bozeman, B., Dietz, J. S., & Gaughan, M. (2001). Scientific and technical human capital: An alternative model for research evaluation. *International Journal of Technology Management*, 22, 716–740. https://doi.org/10.1504/IJTM.2001.002988

Bratt, S., Hemsley, J., Qin, J. & Costa, M. (2017), Big data, big metadata and quantitative study of science: A workflow model for big scientometrics. Proc. Assoc. Info. Sci. Tech., 54: 36–45. doi:10.1002/pra2.2017.14505401005

Costa, M., Qin, J., & Bratt, S. (2016). Emergence of collaboration networks around large-scale data repositories: A study of the genomics community using GenBank. *Scientometrics*, 108(1): 21-40. https://doi.org/10.1007/s11192-016-1954-x.

Crane, D. (1972). *Invisible colleges: Diffusion of Knowledge in Scientific Communities*. Chicago: The University of Chicago Press.

Darko, A., Chan, A. P.C., Adabre, M.A., Edwards, D.J., Hosseini, M.R., Ameyaw, E.E. (2020). Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities. *Automation in Construction*, 112, 103081, https://doi.org/10.1016/j.autcon.2020.10308.

Deelman, E., Gannon, D., Shields, M., & Taylor, I. (2009). Workflows and e-Science: An Overview of Workflow System Features and Capabilities. *Future Generation Computer Systems*, 25(5), 528–540. https://doi.org/10.1016/j.future.2008.06.012.

Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen A.M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.L. (2018). Science of science. *Science*, Mar 2; 359(6379): https://doi.org/10.1126/science.aao0185. PMID: 29496846.

Hemsley, J., Qin, J., Bratt, S., & Smith, A. (2022). Collaboration Networks and Career Trajectories: What Do Metadata from Data Repositories Tell Us? In: *Proceedings of 85th ASIST Annual Meeting, October 28-November 1, 2022, Pittsburgh, PA.*

Hemsley, J., Qin, J., & Bratt, S. (2020). Data to knowledge in action: A longitudinal analysis of GenBank metadata. In: *Proc. Assoc. Info. Sci. Tech.*, https://doi.org/10.1002/pra2.253.

Hertzel, D.H. (2018). Bibliometric research: History [ELIS Classic]. In: *Encyclopedia of Library and Information Sciences.* 4th ed. Boca Raton, FL: CRC Press.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A*,102(46):16569-72. https://doi.org/10.1073/pnas.0507655102.

Hou, J., Yang, X. & Chen, C. Emerging trends and new developments in information science: a document co-citation analysis (2009–2016). *Scientometrics* 115, 869–892 (2018). https://doi.org/10.1007/s11192-018-2695-9

Kumar, S. (2015). Co-authorship networks: a review of the literature. Aslib Journal of Information Management, 67(1): 55-73. https://doi.org/10.1108/AJIM-09-2014-0116.

Li, H., Zhu, Y., & Niu, Y. (2022). Contact tracing research: A literature review based on scientific collaboration network. *International Journal of Environmental Research and Public Health*, 19(15), 9311. MDPI AG. http://dx.doi.org/10.3390/ijerph19159311.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16 (12): 317–324. https://www.jstor.org/stable/24529203.

MacRoberts, M.H. & MacRoberts, B.R. (1989). Problems of citation analysis: A critical review. Journal of the American Society for Information Science, 40(5): 342-349. https://doi.org /10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U.

Martyn, J. (1975). Citation analysis. *Journal of Documentation*, 31(4): 290–297.

Meadows, A. J. (1998). *Communicating Research*. Academic Press a division of Harcourt Brace & Company.

Morone, P. & Taylor, R. (2010). *Knowledge Diffusion and Innovation: Modeling complex Entrepreneurial Behaviors.* Cheltenham, UK: Edward Elgar.

Newman, M. E. J., Barabási, A., Watts, D. J. (2006). The Structure and Dynamics of Networks. United Kingdom: Princeton University Press.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of National Academy of Science*, 98(2): 404-409.

Price, D. J. (1986). Little Science, Big Science... and Beyond (Vol. 480). New York: Columbia University Press.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? Journal of Documentation, 25(4): 348-349.

Qin, J., Hemsley, J., & Bratt, S. (2022). The structural shift and collaboration capacity in GenBank networks: A longitudinal study. *Quantitative Science Study, 1-20*. DOI: https://doi.org/10.1162/qss_a_00181; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9012484/

Qin, J., J. Hemsley, & S. Bratt. (2018). Collaboration capacity: Measuring the impact of cyberinfrastructure-enabled collaboration networks. *Science of Team Science (SCITS) 2018 Conference, Galveston, Texas, May 21-24, 2018.* https://experts.syr.edu/en/publications/collaboration-capacity-measuring-the-impact-of-cyberinfrastructur

Qin, J., Chen, C., Hemsley, J., Greenberg, J., & Wolfram, D. (2018). Big metadata analytics: setting a research Agenda for data-intensive future (BMA2018): workshop at the Association for

Information Science and Technology Annual Meeting, November 14, 2018, Vancouver, Canada. http://metadataetc.org/BMA2018/bma2018.html

Qin, J. (2010). Empirically assessing impact of scholarly research. In: *Proceedings of the iConference, February 3-6, 2010, Champaign, Illinois*. https://www.ideals.illinois.edu/bitstream/handle/2142/14924/qin.pdf?sequence=2

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics*, 11(3), 841-854. https://doi.org/10.1016/j.joi.2017.07.003

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T., & Karsch-Mizrachi, I. (2021). GenBank. Nucleic Acids Research, 49(D1), D92-D96. https://doi.org/10.1093/nar/gkaa1023

Schrader, A. M. (1981). Teaching bibliometrics. *Library Trends*, 30(6), 151.

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association of Information Science and Technology*, 69(1): 6-20. https://doi.org/10.1002/asi.23917.

Summers, E.G. (1983). Bradford's Law and the retrieval of Reading Research Journal literature. *Reading Research Quarterly*, 19(1): 102-109. https://doi.org/10.2307/747340

Wang, D., & Barabási, A. (2021). *The Science of Science*. Cambridge: Cambridge University Press. https://doi.org/0.1017/9781108610834.

Zeng, Y., Wu, L., Bratt, S., & Acuna, D. E. (2020). Assigning credit to scientific datasets using article citation networks. *Journal of Informetrics*, 14(2), https://doi.org/10.1016/j.joi.2020.101013.