Big Data, Big Metadata, and Quantitative Study of Science: A Workflow Model for Big Scientometrics

PAPERS AND POSTERS: LEAVE BLANK proposal
First Author Name
Affiliation
Address
e-mail address

PAPERS AND POSTERS: LEAVE BLANK proposal
Second Author Name
Affiliation
Address
e-mail address

ABSTRACT

Large cyberinfrastructure-enabled data repositories generate massive amounts of metadata, enabling big data analytics to leverage on the intersection of technological and methodological advances in data science for the quantitative study of science. This paper introduces a definition of *big metadata* in the context of scientific data repositories and discusses the challenges in big metadata analytics due to the messiness, lack of structures suitable for analytics, and heterogeneity in such big metadata. A methodological framework is proposed, which contains conceptual and computational workflows intercepting through collaborative documentation. The workflow-based methodological framework promotes transparency and contributes to research reproducibility. The paper also describes the experience and lessons learned from a four-year big metadata project involving all aspects of the workflow-based methodologies. The methodological framework presented in this paper is a timely contribution to the field of scientometrics and the science of science and policy as the potential value of big metadata is drawing more attention from research and policy maker communities.

Keywords
Big metadata analytics, workflows, methodology, scientometrics

1. INTRODUCTION

Research data repositories store massive amounts of data. An example is the GenBank data repository administered by the National Center for Biotechnology Information (NCBI) that stores genetic sequence data and serves "more than 30 terabytes of biomedical data to more than 3.3 million users every day" (NLM, 2015). Every dataset in these repositories includes metadata that documents items such as the intellectual contributors, property rights, knowledge indicators, and many other fields. Thus, since each dataset includes its own metadata, we now have not only big scientific data but also *big metadata*. The GenBank data repository hosted at NCBI, for example, contains 1.9 trillion sequences, with an annual increase rate of 36.8% in 2016, each one of which includes metadata (Benson et al., 2017).

As scientific data is pivotal to modern science, big metadata about scientific data is also pivotal for data discovery, sharing, reuse, crediting authors and research reproducibility. More recently, and with the emergence of data science, such metadata is being used as trace data that offers insight and policy prescriptions about the scientific enterprise. Triangulating these quantitative metadata with social, political, and economic events creates unprecedented opportunities for science of science and policy research in the big data era.

While opportunities for doing research with such vast metadata resources are exciting and promising, there are major challenges in using it to conduct research. The large scale means most work requires computational efforts requiring skilled workers and costly resources. Additionally, the metadata are often messy, complex, ambiguous, and unstructured. Further, idiosyncratic structures and formats in data repositories vary greatly from discipline to discipline. Together, these factors mean that researchers tend to develop one-off custom solutions to process and prepare the data for analysis. The lack of methodological guidance can leave researchers in a constant exploratory data-processing mode. We contend that most big metadata projects will have to execute similar tasks in roughly the same order. Is there a methodological framework that can guide the data processing and analysis in using big metadata from scientific data repositories as the data source? What would such a methodological framework look like?

This paper attempts to address the above research questions through a project case. We argue that the development of workflow methods is a necessary step to increase transparency and contribute to research reproducibility in scientometric work and data-driven policy development, especially when work is iterative, ad hoc, exploratory, and messy. Such a methodology could also help researchers avoid reinventing the wheel for each project and give

ASIST 2017,. Crystal City, VA | Oct. 27-Nov 1, 2017

researchers a descriptive shortcut by allowing them to indicate workflows by name, instead of describing their isolated parts.

To accomplish these goals, we first provide a definition of the term *big metadata* in the context of science data repositories as well as the challenges in using it. Following this is a brief description of the workflow-based methodological framework, with examples from our four-year long big metadata analytics project. The description further illustrated by the GenBank metadata analytics project and lessons learned from our experience. Finally, the paper concludes with a discussion of all aspects of the workflow-based methodologies and future study. The methodological framework presented in this paper is a timely contribution to the field of scientometrics and the science of science and policy as the potential value of big metadata is drawing more attention from research and policy maker communities.

2. BIG METADATA AND THE CHALLENGES

Metadata in scientific data repositories are used to document the creators and attributes of data, record associated projects and physical artifacts that have been used to produce the data, and link to publication output. Some scientific data repositories were established before the web was invented, such as GenBank, while others are relative newcomers, e.g., Dryad (http://datadryad.org) and the Knowledge Network for Biodiversity (KNB), (https://knb.ecoinformatics.org/). Many of these data repositories are international in scope, that is, data may be submitted by anyone in the research community from anywhere.

As researchers submit science data into the repository, they also submit information about the data (metadata). The quality of this metadata depends on the input tool and underlying supporting technologies. Technologies and "x-informatics," as Jim Gray (Hey, Tansley, & Tolle, 2009) famously put it, where x may be substituted by any discipline name (e.g., biomedical and ecological), have supported massive data in physics, astrophysics, biological sciences, ecological systems, geography, and many other fields during the last three decades. The x-informatics phenomenon implies that the tools and technologies were developed in consultation with field specialists who are far more interested in the quality of the scientific data than the metadata that describes it. Thus, across the fields, big metadata are often no less massive or complex than the data they represent.

The term *big metadata* can be found in literature as early as 2013, but it was used to refer to the management of metadata for big data (Smith et al., 2014). Broadly speaking, indexing databases, such as the Web of Science, and catalogs such as WorldCat are in the big metadata domain, but they are not the subject of this paper. In the context of this paper, we use the term *big metadata* to refer to the structured, semi-structured, or unstructured descriptions of scientific data stored in repositories. Such descriptions usually follow a metadata standard recognized by a research community and are created by researchers. For example, ecosystems data in the KNB repository are described by the Ecological Metadata Language (EML) and the tool Morpho[1] was developed to support the use of this community metadata standard in submitting datasets to the repository.

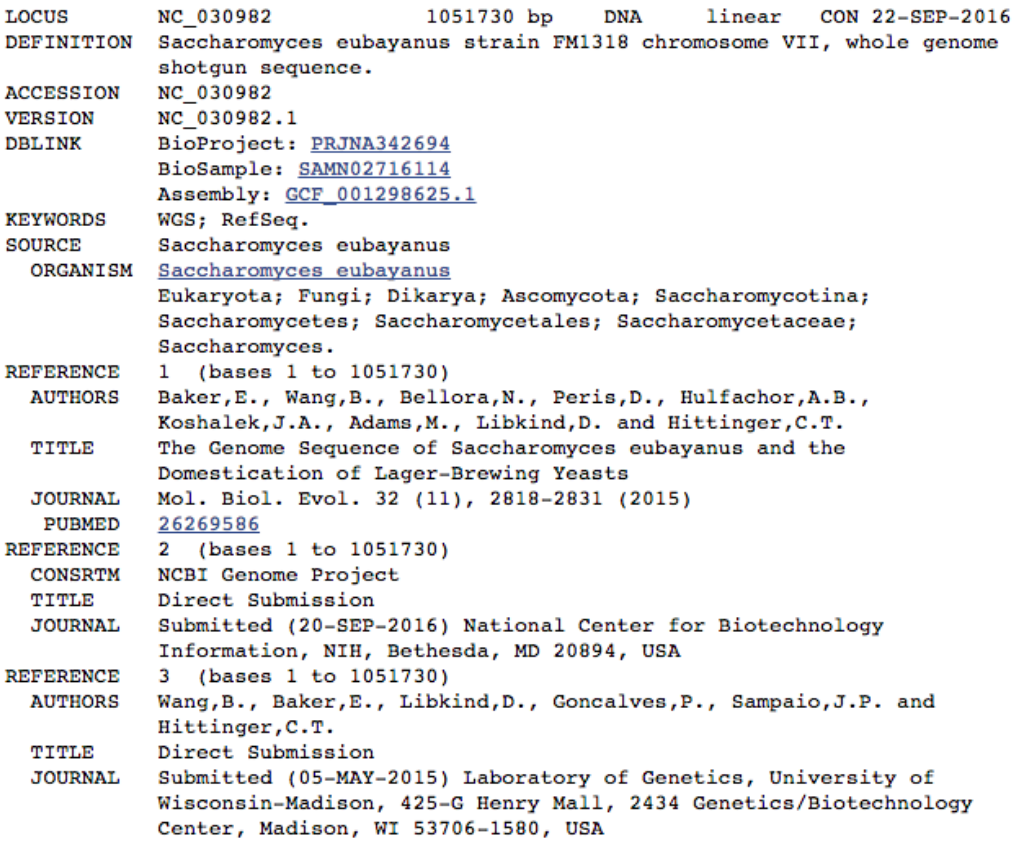
The metadata in such systems can be treated as trace data that supports research exploring a range of questions about the scientific enterprise. For example, we can examine how collaborations and research networks rise and fall, or how the scale and structure of these networks relate to productivity. Linking to additional data sources opens up a vast set of additional questions. We might wish to study the collaborative processes from data generation to the publication of scholarly papers to applying for and securing patents or funding. Alternately, we might analyze the diffusion of knowledge across media or the economic impact of patents or successful grant applications. In other words, these metadata support research in areas with profound social, economic, and policy implications.

The promises of big metadata, however, come with equally big, if not greater, challenges. These challenges require both computationally intensive processing and, more importantly, careful research design. One key computational problem is that names and organizations often need to be disambiguated. Figure 1 provides an example of GenBank metadata record. Personal names include the full last name with first and middle name initials. In big metadata, we find that multiple people can have identical last names and first (and middle) initials. Such ambiguities can damage the accuracy of data and cause negative ripple effects, e.g., to the validity of analytics. We also find inconsistencies and errors in names; in one record a person may list their middle initial and in another they may not. Computationally, such cases will appear to be two different people. Similarly, the same organization may be listed with inconsistent abbreviations, organization units in different orders, or the inclusion or exclusion geographic locations across records. In sum, these concerns can jeopardize the quality of the final analysis. Although standard identifier systems that resolve these ambiguities have begun deployment (e.g. Microsoft's Academic Graph (Sinha et al., 2015)), most older scientific data repositories have not implemented similar systems on new data entry, much less for those in the vast legacy metadata in these repositories. Thus, each new research project must disambiguate named entities, a non-trivial task that makes data processing costly in time and effort. It also makes linking data fields on names, both within the same data collection and across different databases, difficult.

An assumption of many new big metadata researchers is that the open availability (e.g. accessible to FTP) of the data means it is structured, formatted, complete, accurate, standardized, and consistent across sources. For reasons discussed above, such assumptions are more illusory than reality. The range of data problems across the numerous systems means that researchers tend to develop their own

[1] https://www.dataone.org/software-tools/morpho

tools that focus on solving the specific data problems they face. The result is that few general, open source tools exist, placing the burden on other researchers to solve the same, or similar, problems with each new research project. When we note that different researchers have different levels of skills and experience we have to conclude that knowledge claims in our field could rest on dramatically different levels of data cleaning quality.



```
LOCUS       NC_030982            1051730 bp    DNA     linear   CON 22-SEP-2016
DEFINITION  Saccharomyces eubayanus strain FM1318 chromosome VII, whole genome
            shotgun sequence.
ACCESSION   NC_030982
VERSION     NC_030982.1
DBLINK      BioProject: PRJNA342694
            BioSample: SAMN02716114
            Assembly: GCF_001298625.1
KEYWORDS    WGS; RefSeq.
SOURCE      Saccharomyces eubayanus
  ORGANISM  Saccharomyces eubayanus
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE   1  (bases 1 to 1051730)
  AUTHORS   Baker,E., Wang,B., Bellora,N., Peris,D., Hulfachor,A.B.,
            Koshalek,J.A., Adams,M., Libkind,D. and Hittinger,C.T.
  TITLE     The Genome Sequence of Saccharomyces eubayanus and the
            Domestication of Lager-Brewing Yeasts
  JOURNAL   Mol. Biol. Evol. 32 (11), 2818-2831 (2015)
   PUBMED   26269586
REFERENCE   2  (bases 1 to 1051730)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (20-SEP-2016) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (bases 1 to 1051730)
  AUTHORS   Wang,B., Baker,E., Libkind,D., Goncalves,P., Sampaio,J.P. and
            Hittinger,C.T.
  TITLE     Direct Submission
  JOURNAL   Submitted (05-MAY-2015) Laboratory of Genetics, University of
            Wisconsin-Madison, 425-G Henry Mall, 2434 Genetics/Biotechnology
            Center, Madison, WI 53706-1580, USA
```

Figure 1. A sample metadata record for a genetic sequence in GenBank repository (Source: https://www.ncbi.nlm.nih.gov/nuccore/1069890484?report=genbank).

Clearly, the challenges in using big metadata for analytics can lead to pitfalls and affect the validity and reliability of research. Sound methodological approaches and reporting become critical in mitigating some of these challenges. The following section will discuss the methodological approaches to big metadata analytics from a workflow perspective.

3. METHODOLOGIES

As used by industry, *workflow* describes the orchestrated and repeatable patterns of business activities performed in a sequence of operations (BPM Center of Excellence, 2009). With respect to e-science, we can think of workflow orchestration as "the activity of defining the sequence of tasks needed to manage a business or computational science or engineering process. A workflow is a template of such an orchestration" and has a lifecycle that encompasses four broad aspects: composition, mapping, execution, and provenance (Deelman et al., 2009, p. 529). The concept of workflow orchestration and its four broad aspects for e-science sets the context in which the workflow in big metadata analytics is defined. Although Deelman et al.'s four broad aspects of workflows are used to assess workflow tools for highly computational science, they provide a framework for discussing the methodology for big metadata analytics.

A workflow in big metadata analytics is composed of a sequence of tasks and activities that are performed iteratively to achieve the goal of the workflow. A collection of interrelated and interdependent workflows, together with methods and the rationale for why and how the methods are used, forms the methodology (or template) for conducting big metadata analytics. An important characteristic of this workflow-based methodology is *iterativity*, that is, the tasks and activities in workflows are often recurrent, overlapping, and repeated. With every iteration, the methodology is refined and enhanced.

3.1 Workflows in Using Big Metadata

Big metadata analytics consists of several stages that are iterative and overlapping. Two types of workflows are

involved in these stages: conceptual workflows and computational workflows (see Figure 2). Connecting conceptual and computational workflows, and running simultaneously with all stages, is *collaborative documentation*, which is a team-based effort to record goals, rationales, strategies, steps, and activities for search, reuse, and informational purposes. Conceptual workflows identify goals for both the milestones of the study and activity areas for achieving the milestones. At this conceptual level, the rationale and methods will also be articulated and documented. These may be revised and updated iteratively as the research progresses. Corresponding to the conceptual workflows are the computational workflows that essentially codify what has been defined at the conceptual level for machine execution.

Computational workflows may be divided into interdependent task-level and activity-level workflows. Task-level workflows are generally characterized by inputs, outputs, and processes/parameters, where outputs are task specific, e.g., creating code to parse data fields in a table by indicating where to fetch data from, the criteria or conditions used in the parsing, and where the output should be sent to. Task-level workflows tend to be iteratively tested and adjusted, and the successful completion of a task often affects the execution of other task-level workflows. Activity-level workflows have a broader goal than task-level workflows and usually take a group of task-level workflows to accomplish. For example, author name disambiguation is a typical activity in big metadata analytics that requires several steps or processes to complete, depending on the disambiguation strategies employed. Each of the steps or processes involve planning, organizing, and executing some task-level workflows. For big metadata analytic projects, these workflows can be described as analytical pipelines that orchestrate tasks and activities for big metadata analytics projects.

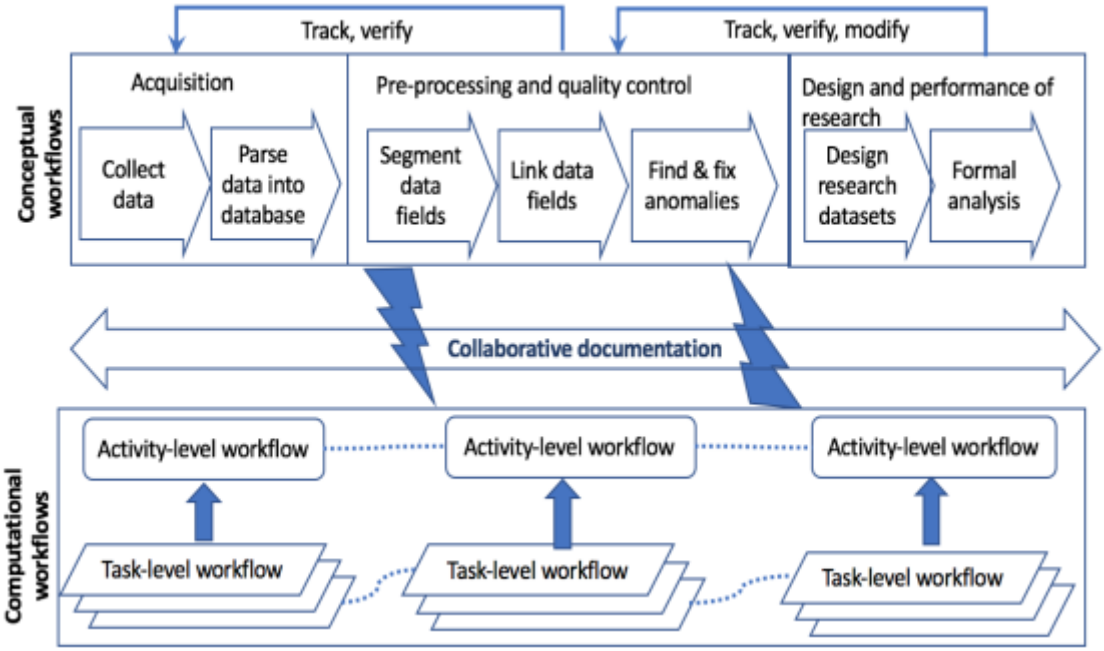


Figure 2. A workflow-based methodological framework for big metadata analytics. The conceptual workflows define the goals and strategies of activities, whereas computational workflows codify the conceptual workflows for machine execution. Any of the stages in a conceptual workflow may be iterated when processing and/or analytic results need to be tracked and verified and methods need to be modified.

The workflow stages in Figure 2 approximately correspond to Stanton's model of the data science areas and lifecycle: *architecture*, *acquisition*, *analysis*, and *archiving* (Stanton, 2013). For our purposes, these data science areas can be refined and interpreted to better reflect the context of big metadata from scientific repositories (see table 1). For big metadata, *acquisition* consists of data collection activities, such as extracting metadata from data repositories and any subsequent data additions from other sources. *Pre-processing and quality control* encompasses computational and other types of data processing (e.g. data cleaning or wrangling), including algorithm design, and script development. It also includes exploratory analysis such as descriptive statistics and visualizations as well as collaborative logical interpretation and synthesis of results of those analyses. *Design and performance of research* refers to the creation of datasets based on research questions and/or any data products.

Table 1. Comparison of areas between big metadata analytics and data science

Areas in Big Metadata Analytics	Areas in Data Science
Acquisition	Acquisition
Pre-processing and quality control	Analysis, Architecture
Design and performance of research	Archiving, Analysis, Architecture
Collaborative documentation	Archiving, Analysis

3.2 Stages and Workflows

3.2.1 Acquire Data

In this stage, decisions are made about the scope and type of metadata to collect based on the intended research, making data collection a part of the conceptual workflow. However, the nature of scientometric analysis often is characterized by uncertainty about what data will best support the research. For example, are we interested in scientific collaboration as the focus of our research? Do we need the scientific data associated with the metadata because understanding its size and complexity may be related to the amount of work done by the submitting actor in a given network? These questions require a consideration of human and technical resource constraints and suggest possible task-level methodological and technical solutions. Thus, they play a central role in this very first step of the big metadata workflow. Ideally, during their deliberations, the team should also consider the "downstream" steps of the workflow, e.g., data processing, data analysis, and interpretation of results.

In our example, the GenBank FTP server stores semi-structured, compressed genetic sequence data files submitted by researchers from around the world. Most of these compressed files are anywhere from 2MB-80MB. As of 8/16/2012 when we downloaded the data, it contained 1,723 data files and 129 index files, all in the compressed .gz format. The size of the compressed sequence data was approximately 75GB and that of the index files 8GB. Each GenBank record contains both metadata describing the submission as well as the genetic sequencing information. The sequencing information comprises the bulk of the file size (over 90% of the file in many cases) and is not needed for the purposes of our study.

3.2.2 Parse data into database

The second step of the workflow is parsing the data into the database. This work often involves data exploration efforts aimed at understanding what the data look like. The challenge here is that the metadata is not structured and often inconsistent. Thus, this step often involves a great deal of data cleaning and checking one's own efforts. Parsing scripts may be developed alongside scripts that test the assumptions of the parsing scripts. We often find that the first pass of the work does not correctly parse all of the data and that additional parsing rules need to be incorporated. Sometimes successive rounds resolve increasingly smaller numbers of errors until we have all, or nearly all, problems resolved. The team collectively evaluates the errors and error rate to determine if there is an acceptable error rate, and if, what that should be. The very process of scripting parsers and resolving errors leads to a deeper understanding of data. We also note that the process is much easier when at least one member of the team has domain knowledge and understands what the data should look like.

The architecture part of this second stage consists of the design and selection of a big metadata analytics infrastructure. Our project's architecture is comprised of an ecosystem of servers, applications, and databases. The data ingestion and storage layer includes a Linux server with a MySQL database. We use Microsoft R Open[2] on an Ubuntu virtual machine for analysis and visualization. R Open includes a refactorization of some base functions to support native parallelization of those functions, meaning that the Microsoft version of R can use the machine's multiple processors, and so is suitable for big metadata analytics. Packages that we have found useful include foreach[3], doParallel[4], data.table[5], Matrix[6], and igraph[7]. The server is configured to support data sharing and multiple concurrent users. Other collaboration technologies that orchestrate the team's work include GitHub, Google Drive, Evernote, Email, and Dropbox.

3.2.3 Segment data fields

Segmenting data in certain fields becomes an unescapable step for almost any big metadata analytic project due to the characteristics of big metadata discussed in the previous section. The need for segmentation arises when a field contains multiple values, each of which needs to be stored separately. Typical data fields that have such conditions include author name field, subject and/or keyword field, author affiliation field, among others. Effective analytics requires a fine granularity and structure in the datasets that are designed to address research questions. Breaking down coarsely granular data fields into smallest data atoms is a step to meet this requirement.

This step often requires some analysis, such as whether the record is flat or hierarchical (nested), and the general

[2] https://mran.revolutionanalytics.com/open/

[3] https://cran.r-project.org/web/packages/foreach/

[4] https://cran.r-project.org/web/packages/doParallel/

[5] https://cran.r-project.org/web/packages/data.table/

[6] https://cran.r-project.org/web/packages/Matrix/

[7] https://cran.r-project.org/web/packages/igraph/

structural complexity. The author name field, for example, often requires several steps to separate individual names into a single name per row. This formatting task may take several iterations to complete because of inconsistently abbreviated names or the inclusion of "and" or "&". More often than not, specific data cleaning techniques such as author name disambiguation have to be employed to get clean, accurate data. Similar to the previous step, segmenting also requires quality checks to verify data assumptions. Thus, we often develop test scripts alongside parsing scripts. As before, working with the data in this way leads to growing depth of knowledge about the data.

3.2.4 Link data fields

In this step, segmented data fields need to be meaningfully connected in the database to prepare for queries that generate required datasets. If the original format of the metadata description was designed for human viewing and understanding, then determining relationship rules will require re-modeling to support big metadata analytics, which requires the data to be meaningfully structured in ways that make computational processing and execution possible. While there are many possible linkages between data fields, we have to be aware of the existing relationships and stay focused on the ones that can help us address the research questions. Questions must be asked such as: What relationships exist in the raw metadata? How might they be reflected in the new database? What new relationships need to be established for addressing the research questions? Linking also requires understanding the data at depth and the nature of the fields to make valid linkages between data tables.

3.2.5 Find and fix data anomalies

Data anomalies can occur because of human error, irregularities or inconsistencies in formatting, or even as computational processing errors by the research team. That is, anomalies can be produced by human error and machine error. In either case, it is important to explore the root of the problem: whether the anomaly was caused by human error or machine error, the error's location in the data lineage, and how it might have happened. Research may be necessary to track down to the originally collected big metadata and use other references to provide clues for solving the anomalies.

Explorative data analysis (Tukey, 1977) at this stage can be useful to discover the shape, structure, and inclusivity of the data to make an informed decision of how to ameliorate the anomalies, and to determine how they will impact our statistical analyses or visualizations. At the task level, descriptive statistics and visualizations can be an effective method to test assumptions about what the data look like. It is also useful to draw subsets of the data out to test the assumptions about the data to ensure the segmenting task accomplished what they were designed to do before applying the code to the whole dataset.

3.2.6 Design datasets based on research questions and hypotheses

Once the data cleaning and quality verification is done, datasets will be extracted based on criteria driven by research questions and hypotheses. In order for the study to stand on a solid methodological ground, the design of the datasets requires a team effort. During this stage, incorporating data from other sources may occur depending on the research plan. If the source needed is not in the plan, then the workflow goes back to the drawing board in first stage. Even without additional data, we do many tests and trials before settling down on the final datasets because the reliability and validity of research findings rely on these data. Also, the final datasets will be shared with the research community and undergo the scrutiny of peer review.

3.2.7 Conduct analysis

Many analytic methods are available for analyzing the datasets: statistical methods, social network analysis, machine learning, visualization, and combinations thereof. From a methodological perspective, the key here is not the specific methods used but rather that the rationale and procedures are well thought through and documented. The importance of a conceptual workflow becomes evident in that it provides a robust template for a well-justified and supported study.

While the big metadata workflow steps are ordinal in their numbering, the stages are often iterative and often lead to more questions or unexpected results. For example, the collection of data precedes parsing and processing, though later we may later realize we need additional data, and segmenting data fields could be revisited as more is learned about the data. Thus, the workflow is conceived as an executable flow of stages not because they are in a rigid chronology but because they unfold in a roughly stepwise manner.

In its entirety, the process is about developing a deeper and deeper understanding of the data: what is in it, what its limitations are, what structures and patterns there are in the data, and given its quality, what kinds of questions the data may support answering. The exploratory steps of data verification, question surfacing, and examining data structures are central to a workflow of big metadata when a lack of shared standards, documentation, and structure make uncertainty a staple of big metadata analytics. A key strength of a named workflow, then, is that it facilitates organization and determination of where to return in the workflow during iteration.

3.3 Cross-Stage Components

3.3.1 Collaborative Documentation

Documentation occurs, and often is required, at almost every stage of a research lifecycle and throughout both conceptual and computational workflows. Documentation is done for different purposes and with various tools. The most striking characteristic in team projects is the collaboration involved in creating, maintaining, and managing such documentation. Collaboration in documentation is guided by the data

management plan and orchestrated by clear policy and persistent enforcement within a team.

Collaborative documentation becomes critically important when task-based and activity-based workflows are created by different team members, as well as when there is a change of team member composition. The choreography of the artifacts generated from collaborative documentation serves as the cohesive medium for maintaining coherence between distributed team members' analyses and results. That is, the processes are held together by making sense of graphical results, and writing down findings, issues, and the logic of the steps they took in analysis. In this stage, the research team produces documents such as memos and notes with tools such as Evernote and cloud storage systems to record and share their work.

3.3.2 Data wrangling

Data wrangling consists of any transformations needed to support exploratory analysis, quality checks, or final analyses. Aggregations, filtering, sub-setting, and sorting are all data wrangling tasks. Sometimes called data cleaning, data wrangling happens at multiple points during the workflow and can be, by itself, an iterative operation. Exploratory data analysis (EDA) (Tukey 1977) is a crucial element of this work. We use it to understand how the data is structured before and after any data transformations.

3.3.3 Data Quality Checks

Data quality checks include any efforts to ensure the data appear as expected. Different from finding and fixing anomalies, the focus of quality checks is on verifying the results of script runs and data transformations. For example, once we insert data into the database we run checks to ensure it was inserted correctly. We note here that we embed in the code our assumptions about how it does and should work. However, because our assumptions are not always right, part of a quality check is thinking through assumptions. Team discussions help uncover these hidden assumptions and help identify what should be tested and how to test it.

4. THE CASE OF GENBANK

In this section, we illustrate the use of workflow-based methodology in our big metadata analytics project and demonstrate how the stages and workflows iteratively unfold as research commonly progresses. The goal of this project was to demonstrate the feasibility of using big metadata for studying the impact of cyberinfrastructure-enabled science on the formation, thriving, and cooling down of research collaboration networks over time. Collaboration networks that arise from authorship of data and publications may be considered a symbolic representation of scientific capacity of a country or institute. The project went through the workflow stages from data acquisition through to analysis (Section 3.2.1-7) after numerous iterations. The final data collection includes not only the complete metadata records from GenBank made available at the time of downloading from the FTP site, but also the NCBI Taxonomy[8] that was incorporated into the database and linked to GenBank metadata records based on the taxon lineage in the records. The incorporation of NCBI Taxonomy makes it possible for big metadata analytics based on taxonomic classes.

In the big metadata collection from the GenBank FTP server, we linked segmented attributes by identifying relationship rules, for instance, 1) a publication metadata record is associated with one or many authors; 2) a data submission is associated with one or many annotations; and 3) an annotation is associated with one or many data submissions. From these rules, it is clear that a many-to-many relationship between the annotations and data submission records requires linking to normalize the fields.

To test a series of workflows for studying a sub-community in the GenBank data repository, we selected metadata records for DNA sequences related to the West Nile Virus because outbreaks in infectious diseases often generate new discoveries of virus genes that result in a concentrated period and a higher frequency of data submissions. In designing the conceptual workflows, the team collected information from the Center for Disease Control (CDC) to learn about the history and extent of the disease's spread. Studying this research community through big metadata can help identify the hub of researchers and new discoveries of virus strains, and possibly link the basic research to clinical and drug developments.

The first step was to extract data by querying our database. Because the study of the West Nile Virus research community required we use a subset of our preprocessed data, certain stages related to data parsing and cleaning were performed only when special needs arose. The subset of data included records of publications, data submissions, and patents associated with West Nile Virus in our database. In the conceptual workflows, several key areas helped the computational workflows stay on track: 1) we developed a concept map, or copy of the database schema, showing the relationships of the records and indicating which metadata were necessary and sufficient to ensure validity; 2) we leveraged background knowledge in the biological sciences to ascertain that the query was representative of the object intended for extraction; and 3) we checked the internal logic of the algorithm, as well as the meaning of the metadata, to validate that our query was consistent, valid, and thorough.

For example, in executing a query for the West Nile Virus records, the analyst must know that a) the default setting of the regular expression matching algorithm is "greedy," and returns both upper and lower case entries and b) the "title" field includes data submissions. After meeting with a more senior project researcher, the analyst received feedback on her work and found that the grep() algorithm *did not* match

[8] The NCBI Taxonomy database is a curated classification and nomenclature for all of the organisms in the public sequence database. https://www.ncbi.nlm.nih.gov/taxonomy

the data submission for West Nile; for these records, the analyst must join across database tables to query for the precise taxonomy IDs that corresponds to West Nile Virus. As we proceeded with finding and fixing anomalies in this way, we discovered that there were twelve taxonomy IDs describing West Nile Virus data submissions, which were best suited for the research questions. Another example arose when we realized that if science policy prioritized West Nile Virus *environmental* research during the time we chose, then the data extraction method should include the taxonomy ID that reflects 'ENV' (i.e., environmental sampling) sequences.

The formal research was performed on the extracted subset of data to explore the network patterns, research productivity, and knowledge diffusion within the West Nile Virus authors, inventors, and scientists. Since the patent data in GenBank was a new addition to the whole data collection and the patent inventors' names were not previously disambiguated, the research team went back to the parsing and preprocessing stages to parse, segment, and disambiguate the inventor name by utilizing algorithms and dataset from Li et al. (2011) and Lai et al. (2011). Next, we visualized the network and analyzed network and node-level statistical properties to gain a better sense of the modularity and community structures such as transitivity, and clustering coefficients. We also identified node-level properties such as the distribution of centrality scores for individual nodes, with relations defined as co-authorship, co-sequence submission, or co-patent invention (Costa, 2016).

These exploratory and confirmatory analyses were extended as we dug deeper into the West Nile Virus metadata to compare it with other infectious disease research communities, such as SARS, H1N1, and Influenza. Sporadic data wrangling and quality checks were conducted along the way as needed. Further, we found that though many members of the team had moved on, their documentation made the reuse of scripts and quality checks fairly easy. An important activity implied but not explicitly entered in the workflows was that, through the process, we met a number of times a week, communicated via email and shared files in Evernote and Dropbox, and consulted with team members to synthesize results and disseminate significant findings.

4.1 Lessons learned: Documentation, documentation, and documentation

The importance of documentation during a research lifecycle cannot be overemphasized. Procedures, workflows, programming code strategies, research ideas and questions alike should be well documented. The nature of our methodologies required sustained scrutiny over data quality. Documentation offers a valuable repertoire for provenance of research information.

Documentation contributes to the transparency and reproducibility of research. Using collaboration tools to collaboratively document methodologies used at different stages, especially where iterations took place, can greatly enhance the completeness and accuracy of such documentation. For this project, our team used the following tools for collaborative documentation, management, versioning, and group sharing.

GitHub: Processing documentation, documented workflows and code are stored here, which provides history and version tracking.

Evernote: On-the-fly naming for each message, with datasets, scripts, and other artifacts from analysis are attached in notes.

RStudio: A dedicated server on which team members have accounts where formal analysis and collaborative documentation is done.

Email: Notes from email discussing technical concerns, such as the connection to the database, memory allocation issues opened and resolved with IT personnel, all these issues are integral to the workflow documentation, and data provenance.

MySQL Workbench: Justifying steps in comments and stored procedures and extracting research datasets comprise the use of MySQL. Referencing schema such as the Entity Relationship Diagram (ERD) also helped to inform the researchers' extraction of datasets.

Dropbox: Used for file storage and document sharing, and as a citation repository before the introduction of a citation manager such as Zotero. Overwriting and co-editing documents presents issues, but it is used for file and R script sharing.

Google Drive: Documents such as notes and papers reporting research findings are shared among team members to enable collaborative writing and editing. Files are exchanged here when space constraints in other platforms proved limiting.

Drupal: The public face of the project is maintained through the content management system, which serves as the backend of the project website. Dissemination of research papers, data products, and other media are housed on the site which is managed by a subset of team members.

It is worth noting that documenting all steps involved in big metadata analytics is impractical nor necessary. Information critical to methodology justification, reproducibility, and continuity must be documented formally by an Evernote entry or a text document; whereas small adjustments of code in computational workflow creation and testing are not needed since GitHub tracks changes and we can comment the work. Automated documentation aids the iterative process of data analysis and reporting immensely; however, this tends to be available only for simple descriptions, such as keyword searching old Evernote documents, email messages, Dropbox files, GitHub commits, R scripts. The search algorithms for each are also different, requiring specialized knowledge for finding old files, notes, or data objects. Yet such automated and manual documentation is critical for revisiting analyses over time.

4.2 Lessons learned: *Know Thy Methodology, Know Thy Data*

Big metadata analytics requires careful design of datasets based on research questions. A misstep in data prep may cause a stalled server, never-ending loops, or very large datasets being exploded out of proportion. When we initially started working with GenBank data, the massive size of 24 million rows caused server stalls due to excessive memory usage, and none of the team members could access the working environment.

In cases of not-so-huge data, we also encountered cases where data caused issues like disconnection from the network due to inactivity for a long time. The main reason behind this challenge was multiple or never ending loops. A classic example of one such scenario is when SQL queries require multiple tables to be joined to filter out the required data.

There were also instances where, while merging datasets, data exploded out of proportion due to data matching issues. For example, in GenBank metadata, every patent has multiple references. When trying to merge authors who were involved in research pertaining to a certain organism, we decided to merge it with data from Dataverse repository data. Due to multiple references per column the dataset exploded from some thousand rows to tens of millions of rows.

It is also important that we understand if the metadata adheres to any published content standards or best practices, such as the international data repository bioinformatics standards, publication and submission reference standards, or U.S. Patent and Trademark Office (USPTO) standards, and whether the fields use any known and shared vocabularies. Linking data fields pivots on the extent to which metadata conforms to these employed standards, the formal or informal practices of data entry, and vocabularies. For example, in our project we were able to take advantage of the standardized fields and effectively link USPTO data (Li et al., 2014) to NCBI GenBank metadata and NCBI Taxonomy database to create a richer, triangulated big metadata system.

Detecting data anomalies and ameliorating data entry errors generally involves writing code to check the work that has been done. For many analysts, this validity checking is not explicit, but an intuitive part of the data exploration and confirmation process. It can get lost if documentation consists of, e.g., an R script that is not commented at all or contains comments that are specific to the analysis, in language that does not make the thoughts of the analyst transparent.

5. DISCUSSION AND CONCLUSION

Big metadata analytics has a great potential for studying the science enterprise at an unprecedented scale. The proliferation of data and metadata in scientific repositories makes linking and triangulating multiple data sources possible, which promotes the reliability, transparency, and reproducibility of scientometric research. Based on our project experience, we generalized the challenges in using big metadata as a source of data for scientometric analytics and developed a methodological framework for big metadata analytics. The methodological framework was motivated by our own experience in mining one of the largest scientific data repositories as well as the pitfalls we encountered and mistakes we made. Although the methodological framework is created in the context of big metadata analytics, it is also possible that it be applied to other areas of big data analytics where similar data situations are encountered.

To use big metadata (or, big data for that matter) effectively and meaningfully, it is more important to be able to articulate the research question and the methodology used to address the research question then simply calculating statistics and making data visualization. The methodological framework we developed offers guidance on a workflow-based approach to working with big metadata. It adds a novel way of thinking organizing work to address the problems and challenges of working with big metadata.

While previous studies repeatedly emphasize the need for improved metadata management and provide examples of how their laboratory or company faced and addressed issues in big metadata analytics, solutions are often *ad hoc*, one-off fixes without formalized workflows. The *ad hoc* approaches are neither generalizable nor scalable for big metadata analytics. Big metadata analytics performed on an *ad hoc* basis is more likely to suffer from mistakes made in both conceptual and computational workflows. The research may become costly, or worse, result in methodologically fundamental problems that can undermine the reliability and validity of research. Without explicit workflows, metadata analysis remains opaque, irreproducible, *ad hoc*, and inaccessible. To the extent that data-driven science policy relies on accurate, careful, big metadata analysis, our methodological model of workflows fully exploits opportunities offered by the confluence of data repositories and data science. The opportunities highlighted by scientometric researchers are great, yet can are often disparate, one-off analyses. Such work can be guided by data science workflow theory. That is, start from the proposed methodological model to better leverage the intersection of technological and methodological advances in data science and the emergence of large-scale scientific data repositories that generate massive amounts of metadata.

Future work will explore the applicability of our methodological framework for data analytics outside of the scientific data repository domain. For example, in what ways are scientific repository big metadata different from traditionally-conceived big data in terms of science of science and policy research? Likewise, we want to explore what distinguishes big metadata science from its corollary big data science, and compare how workflows in analysis of big data are similar. In summary, the methodological framework presented in this paper is needed contribution to both the scientometrics and science of science and policy as

data repository metadata will remain central to science going forward.

Acknowledgement: The authors thank the support of NSF SciSIP grant #1561348 and Shrutik Katchhi and Suchitra Deekshitula for their technical assistance.

REFERENCES

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, *45*(Database Issue), D37–D42. https://doi.org/http://doi.org/10.1093/nar/gkw1070

BPM Center of Excellence. (2009, October 26). Business Process Management Center of Excellence Glossary [Government]. Retrieved from https://www.ftb.ca.gov/aboutFTB/Projects/ITSP/BPM_Glossary.pdf

Costa, M. R. (2016). *The Interdependence of Scientists in the Era of Team Science: An Exploratory Study Using Temporal Network Analysis*. (Dissertation). Syracuse University, Syracuse, NY. Retrieved from http://surface.syr.edu/etd/425

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

Galison, P., & Hevly, B. W. (1992). *Big science: The growth of large-scale research*. Stanford University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=zGmsRqIILx8C&oi=fnd&pg=PR9&dq=Galison,+P.+1992.+Big+Science:+The+Growth+of+Large-Scale+Research.+Stanford,+CA:+Stanford+University+Press&ots=P_gg5wpg7Y&sig=Zpv-uEVQhIvHbT37G8MfnQi2YmE

Greenberg, J., White, H. C., Carrier, S., & Scherle, R. (2009). A metadata best practice for a scientific data repository. *Journal of Library Metadata*, *9*(3-4), 194–212. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/19386380903405090

Hey, tony, Tansley, S., & Tolle, K. T. (2009). Jim Gray on eScience: A Transformed Scientific Method. In *The Fourth Paradigm: Data Intensive Scientific Discovery* (First Edition, pp. xvii–xxxi). Redmond, WA: Microsoft. Retrieved from http://languagelog.ldc.upenn.edu/myl/JimGrayOnE-Science.pdf

Interactive Crowdsourced Literature Key-Value Annotation Library - iCLiKVAL. (n.d.). Retrieved April 10, 2017, from http://iclikval.riken.jp/

Lai, ronald, D'Amour, A., Yu, A., & Fleming, L. (2011). Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database (1975 - 2010) [Dataset]. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/15705

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., … Fleming, L. (2014). Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). *Research Policy*, *43*(6), 941–955. https://doi.org/http://doi.org/10.1016/j.respol.2014.01.012

Light, R. P., Polley, D. E., & Börner, K. (2014). Open data and open code for big science of science studies. *Scientometrics*, *101*(2), 1535–1551. Retrieved from http://link.springer.com/article/10.1007/s11192-014-1238-2

NLM. (2015). Congressional Justification FY 2015 [Document]. Retrieved from https://www.nlm.nih.gov/about/2015CJ.html

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B. (Paul), & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web*. Florence, Italy: ACM. https://doi.org/10.1145/2740908.2742839

Smith, K., Seligman, L., Rosenthal, A., Kurcz, C., Greer, M., Macheret, C., … Eckstein, A. (2014). "Big Metadata": The Need for Principled Metadata Management in Big Data Ecosystems. In *Proceedings of Workshop on Data Analytics in the Cloud* (pp. 13:1–4). Snowbird, UT, USA: ACM. https://doi.org/10.1145/2627770.2627776

Stanton, J. (2013). *Introduction to Data Science*. (Third Edition, Vols. 1–1). Syracuse, NY: Syracuse University. Retrieved from https://itunes.apple.com/us/book/introduction-to-data-science/id529088127?mt=11

Tukey, J. W. (1977). *Exploratory Data Analysis. Pearson, Reading, Mass*. Reading, Mass.: Pearson.

Weingart, S., Guo, H., Börner, K., Boyack, K. W., Linnemeier, M. W., Duhon, R. J., … Biberstine, J. (2010). Science of Science (Sci2) tool user manual. *Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington. Retrieved January*, *15*, 2011.

Özdemir, V., Kolker, E., Hotez, P. J., Mohin, S., Prainsack, B., Wynne, B., … others. (2014). Ready to put metadata on the post-2015 development agenda? Linking data publications to responsible innovation and science diplomacy. *Omics: A Journal of Integrative Biology*, *18*(1), 1–9. Retrieved from http://online.liebertpub.com/doi/abs/10.1089/omi.2013.0170