Data Accessibility Paper

Joanna Colovas

- Abstract
- Importance
  - Incentivize authors to publish/make available their original data
  - Publishing data helps get more use out of research
  - Helps eliminate file drawer effect as it shows negative data
- Keywords
  - Data accessibility
  - Data reproducibility
- Introduction

  *Need more framing of the problem** - DA upon request isn't good enough

  Scientific Data as a Public Good The United States Government spent over two hundred million dollars (USD) in 2024 on research expenditures (@congress.gov2024). The result of all of these investments are data, paid for in part by taxpayers. Therefore, data is a public good, and best used as a benefit to those who provided the funds for it. Once data has been generated, it can be used not only for initial analyses, but over and over again in future studies or meta analyses. Additionally, data can be used to eliminate possible solutions to a problem by the publishing of "negative data." Thinking of data as a public good, if negative data is published, it can help researchers avoid sinking time and financial resources into the investigation of non-viable hypotheses. This non-fruitful investigation is more commonly known as the "filedrawer effect." Data as a public good also is subject to the tragedy of the commons. If no one contributes to the public "commons", data that is available for public use, how can we advance our understanding of our study systems?

  Current DA Policies Current data availability and reproducibility guidelines have been informed by a number of policies created by funding agencies, peer-review journals, conference and special task groups, as well as community interest groups. In 2011, after the Future of Research Communication (FoRC) conference in Germany to establish FORCE11, a community interest group which seeks to encourage and promote data

availability standards. Also in 2011, the Genomic Standards Consortium (GSC) published the MIMARKS/MIxS standards in *Nature Biotechnology* to promote the publication of the "minimum information about a marker gene sequence" (MIMARKS) or "minimum information about x sequence" (MIxS) (@Mimarks2011). These standards are checklists useable by data generators and uploaders towards inclusion of relevant data with sequence uploads in the International Nucleotide Sequence Database Collaboration (INSDC).

In 2014 the group published the Joint Declaration of Data Citation Principles (JDDCP), a document towards the standardization of data citation and its future avaialability(@Data Citation Synthesis Group 2014). The Findable, Accessible, Interoperable, and Reuseable (FAIR) data science guiding principles were put forth in 2016 by Wilkinson et al in *Nature Scientific Data* urges readers to "improve the infrastructure supporting the reuse of scholarly data" (@Wilkinson 2016). Neither the JDDCP nor the FAIR principles are enforcable by any agency. Finally, the National Institutes of Health (NIH) began enforcing the "Policy for Data Management and Sharing" (NOT-OD-21-013) in January of 2023, requiring NIH funded studies to submit a data management and sharing plan (DMS) with their funding applications, and comply with their DMS plan after generation and publication of the funded work(@NIH2023). Non-compliance with NOT-OD-21-013 is identified by funding agencies during annual Research Performance Progress Reports (RPPRs), and may impact future funding decisions (@NIH2023).

ASM Journals With the advent of next generation sequencing, microbiology reserach has generated large amounts of sequencing data, and it is common to upload sequence data to a public repository as well as to include data in research publications. The American Society for Microbiology(ASM) is the major professional body recognized by microbiologists. They have eighteen journals, thirteen primary research journals, three review journals, and two archive journals. In addition, several journals have been folded into others or renamed over time. The ASM family of journals requires that authors "make data fully available, without restriction, except in rare circumstances" (@ASM open data policy). They have adapted this policy from journals *Microbial Genomics* and *PLOS*. In the ASM open data policy they describe the use of a "Data Availability Statement" which includes "data description, name(s) of the repositories, and digital object identifiers (DOIs) or accession numbers" and encourages publishing data on relevant public repositories (@ASM website open data policy). Consequences of non-compliance to the ASM open data policy include contacting research article authors to inform of non-compliance, publication of an "Expression of Concern" for the author and their compliance issues, sanctions on publication in ASM journals, as well as contacting the affiliated research institution and/or funding agencies of the authors (@ASMcompliance). We endeavor to evaluate how well the microbiology community is using reproducible data practices as we believe that this group of researchers will be early adopters of the technologies available as a result of both the ASM and NIH policies towards data availability.

Nucleic Acid Sequencing Efforts

– ft lauderdale accord Nucleic acid sequencing efforts are commonly uploaded to databases to identify appropriate sequences for comparison. There are three major databases worldwide to support sequencing and sharing efforts. The National Library of Medicine's (NLM) National Center for Biotechnology (NCBI) in the United States, the Research Organization of Information Systems' (ROIS) National Institute of Genetics (NIG) in Japan, and the European Molecular Biology Lab's (EMBL) European Bioinformatics Institue (EBI) in Europe. These three databases are part of the International Nucleotide Sequence Database Collaboration (INSDC). These large databases make research by comparison possible. Genetic lineages of microbes are determined by creating phylogenetic trees which compare a new sequence to existing sequences. Phylogenetic trees show how closely related a new microbial genetic sequence is related to others studied before both in terms of evolution and mutation and in structure and function. An important tool for creating phylogenies is the NCBI Basic Local Alignment Search Tool (BLAST) (@altschul1990). The BLAST algorithm allows users to compare a nucleic acid or protein sequence to the NCBI database of over 1TB of data to find similar and related sequences. Without the upload of seqeuces to the NCBI database, the use and success of BLAST would not be possible, despite the effort to upload of sequences to one of the INSDC databases.

Examples of Data Avalability A key tenent of the scientific method is the ability to replicate sceintific findings to ensure that they are not due to error. One way that scientific findings can be replicated is by re-completing the same analyses by another researcher. This is only possible if the data used to complete the original analyses is available for use. The availability of datasets also allows new questions to be answered with existing data or the combination of multiple datasets, such as the use of the Human Microbiome Project's (HMP) sequencing data by researchers to create over 650 scientific publications (@HMP), and the completion of metadata studies. Availablity of data contributed to the rapid sequening of the SARS-CoV-2 virus during the 2020 pandemic and subsequent expedition of vaccine development (@needCOVIDref).

Need another summary of the study and how and why in this paragraph The objective of this study is to determine the current state of data availability in twelve primary research journals from the ASM family of journals using machine learning models.

- Results General Description of the Experiment We set out to determine the current state of data availablility in twelve primary research journals from the ASM family of journals. This objective was completed by first acquiring all papers published in the journals of interest between 2000 and 2024 using the Crossref database and command line tools. We then trained random forest machine learning models to differentiate if each paper contained "New Sequencing Data" (NSD) and then if the paper had "Data Available" (DA). To avoid overfitting the model, we trained each model multiple times, performing validations on a subset of data after each iteration. This allowed us to have a greater number of papers in the training dataset, as well as to have great accuracy and precision

within our models. Using our trained models, we were able to classify over 150,000 papers from the whole dataset to determine if they were NSD or DA. After this, we could perform statistical modeling to describe the data, and generate summary statistics. We were especially interested in the ways in which NSD and DA impact citation metrics.

ASM Journals We used twelve of the ASM primary research journals in this study. Of note, several journals had changes to their publication goals during the 2000-2024 time period. The *Journal of Bacteriology* was the primary place to publish new genome announcements until 2013 when ASM announced journal *Genome Announcements* as a more permanent place for this type of data. *Genome Announcements* was active from 2013 until 2018, when it was rebranded to *Microbiology Resource Announcements*, which has been active from 2018 until present. These two journals appear separately in our analyses as a result of the Crossref database. New genome annoucements have a high percentage of NSD papers, and a high percentage of DA within those papers. As a result, the *Journal of Bacteriology* has had fewer NSD papers, and fewer papers with DA since 2013. Another journal of note is *Microbiology Spectrum* and its rebrand. From 2013 until the fall of 2021, *Microbiology Spectrum* was a review journal. After this point, *Microbiology Spectrum* became a primary research journal. Review journals are less likely to publish articles with NSD, and to have DA. Several journals, including *Microbiology Spectrum*, do not span the entire time period for the study. Journals *mBio* (b.2010), *Microbiology Spectrum* (b. 2013, rebrand 2021), *mSphere* (b. 2016), *mSystems* (b. 2016), and *Genome Announcements* (2013-2018). Not all journals are equally likely to contain NSD and have sequencing DA as a result of their field of interest. Journals *Antimicrobial Agents and Chemotherapy; Infection and Immunity;* and the* Journal of Microbiology and Biology Education;* are all less likely to contain NSD and have DA than the other journals in the dataset.

Descriptive Statistics:

Whole Dataset Using the Crossref database, with validation from WOS, NCBI, and Scopus databases, we downloaded 155779 unique records of papers published in ASM journals between 2000 and 2024. After webscraping there were YYYY papers downloaded with text available for the models to be applied on, as not all records had text available. See table YYYY for number of papers downloaded from each journal. Overall, XXXX% of papers had NSD, and YYYY% of papers with NSD, had DA. See table YYYY for percentages of NSD and DA for each journal. The journal with the highest rate of NSD was XXXXX at XXXX%, and the lowest was XXXX at XXXX%. The journal with the highest rate of DA was XXXXX at XXXX%, and the lowest was XXXX at XXXX%. This was expected as *Microbiology Resource Annoucements* publishes mainly new genomic sequence data and makes the data available. On average, papers in the dataset had XXXX citations/article. This number varies by journal, see table YYYY for data by journal. The journal with the highest rate of citations/article was XXXXX at XXXX%, and the lowest was XXXX at XXXX%. **SENTENCE ABOUT THIS BEING GOOD AND WHAT WE WANTED TO SEE IN THIS DATASET**.

The journals in the dataset span years 2000-2024. See table YYYYY for the distribution of papers per year in the dataset.

Training Dataset **(pretty much the same as the above paragraph, but with stats for the training dataset)** We created a subset of the whole dataset to train our machine learning models. The training dataset initially had N = 250 papers, but was increased over time due to gaps in the dataset, and validation of the trained models (see below). Overall, XXXX% of papers had NSD, and YYYY% of papers with NSD, had DA. See table YYYY for percentages of NSD and DA for each journal. The journal with the highest rate of NSD was XXXXX at XXXX%, and the lowest was XXXX at XXXX%. The journal with the highest rate of DA was XXXXX at XXXX%, and the lowest was XXXX at XXXX%. This was expected as *Microbiology Resource Annoucements* publishes mainly new genomic sequence data and makes the data available. On average, papers in the dataset had XXXX citations/article. This number varies by journal, see table YYYY for data by journal. The journal with the highest rate of citations/article was XXXXX at XXXX%, and the lowest was XXXX at XXXX%. **SENTENCE ABOUT THIS BEING GOOD AND WHAT WE WANTED TO SEE IN THIS DATASET**. The journals in the dataset span years 2000-2024. See table YYYYY for the distribution of papers per year in the dataset.

Descriptive Statisitcs about the Trained Models Two random forest models were trained to predict if published scientific papers "contained new sequence data" (NSD), and if the paper "had data available" (DA), one model for each variable. Other models such as generalized linear regression and boosted trees were explored, but were ultimately discarded in favor of the random forest model (data not shown). Random forest models were chosen to aid in this classification problem as the creation of many decision trees helps to improve accuracy and precision. This type of model has one hyperparameter, 'mtry' or the number of predictors to be sampled at each decision. During iterative model training, a subset of papers were validated after each completed training and deployment of each model. Papers from each journal, and extras from certain journals were hand-validated aginst model predictions to generate confusion matrices. Confusion matricies for the final version of each trained model are available in YYYY table(supplement?). The NSD model used an mtry value of 300 had an Area Under the Curve(AUC) of 0.9549XXX and an acccuracy of 0.8925XXX. The sensitivity of the NSD model was 0.9130XXX, and the specificity of the model was 0.8683XXX. The DA model used an mtry value of 400 had an AUC of 0.9824XXX and an acccuracy of 0.9353XXX. The sensitivity of the DA model was 0.9545XXX, and the specificity of the model was 0.9005XXX (@Data/final/best_model_stats.csv for internal ref). This shows that the models fit the data well, and can provide classifications on new data with an expected error rate of approxmiately 10%. We deemed this as acceptable, accounting for variability in papers and data, and due to the large size of the dataset to deploy the models on.

Regression Model using Negative Binomial Models In this study we sought to investigate the effect of NSD and DA on the number of citations recieved by a given paper. We

focused on NSD papers to determine the effect of having DA. This led us to the use of a negative binomial regression model to best describe our data. All regression data was NSD yes. We focused on the continuous outcome of "number of citations" with predictor variables journal (categorical), age in months (continuous), and DA status (dichotomous). Due to the number of citations being fairly bell shaped with a long right tail (very few papers at advanced age with many citations), the model that best described our data was the negative binomial regression model. This model includes a dispersion parameter to describe the spread of the data. We applied a log transformation to our age variable (age.in.months) to better describe the relationship between time and number of citations received. ???HOW DO I DESCRIBE THIS MODEL??? In general, we found that NSD papers that made DA received more citations over time than those that did not. See figure YYYY for trends in each major journal. ??HOW else do i describe the figure??

- Discussion
  - Percent of ASM papers that have new sequence data available (nsd Yes, da Yes)
    - trends by journal
    - trends by year
  - Making data available
    - provides more citations per paper than not doing so by *XXX* number
    - Allows for replication of studies
  - Why bother doing this?
    - What advantage does it give you as a data generator
    - Is it worth the work?
  -

- Materials and Methods

  Creation of the Training set To train our random forest machine learning model, we first created an appropriate training data set. Using the crossref database, we first downloaded all papers from the selected ASM family of journals from the time period beginning January 1st, 2000, and ending on December 31st, 2024. The data was updated as of February 10th, 2025 with all citation counts frozen at that date. For our initial training set, we chose N = 500 papers from across each journal and time period, adding special emphasis to include papers that were part of our desired set of interest (i.e. contained published data) to ensure that our two models could adequately characterize each paper as a new sequencing paper and if it published raw sequencing data or not. After creating our initial dataset, it was necessary to identify the status of both variables by hand and determine if each paper contained "new sequencing data" (NSD), and if each one had "data available" (DA). This was completed by opening each paper in an internet browser window, and searching for a "data availability" or similar statement. See table XXX for specific cases and how each of these cases were identified for the purpose of this study.

Adding Additional Training Set Papers After initial trainings of our random forest models, a random sampling of papers was collected for each journal to audit the efficacy of the models. To audit the efficacy of the models, we hand identified the status of both variables of interest, NSD and DA. We looked for weaknesses in the models, and updated methodology to reflect important areas of interest. For example, in 2023 the ASM journals changed their formatting to include the data availability statement of a paper in a sidebar of the webpage. We identified this by noticing that all papers from journal *Microbiology Resource Announcements* from 2023-2024 were incorrectly characterized by the model as DA = No. The sidebar of the webpage was not included in the text the model was considering, and code had to be updated to include all sidebar data for all papers. These improvements to the model created a larger and more comprehensive training set of N = 9XX. These validations allowed us to create confusion matricies for each model. Confusion matricies for the final version of each trained model are available in YYYY table(supplement?).

Descriptive Statistics about the Training Set There were XXX papers in the training data set from 12 ASM journals. These papers came from journals *Applied and Envrionmental Microbiology; Antimicrobial Agents and Chemotherapy; Infection and Immunity; Journal of Clinical Biology; Journal of Virology; Journal of Bacteriology; Journal of Microbiology and Biology Education; Microbiology Resource Announcments* (formerly known as *Genome Announcements*); * mSystems; mSphere; mBio; and Microbiology Spectrum.* See table XXXX for the number of papers from each journal in the training dataset. The training dataset includes journal articles published between January 1st, 2000 and December 31st, 2024. See table XXXXX for the number of papers included from each year from 2000-2024. XXX% of training set data was NSD = Yes, and XXX% of training set data was DA = Yes.

Creation of the Training Data from Training dataset To perform the computational steps required for these experiments, we used the python tool Snakemake (@snakeref), and the University of Michigan's high performance computing cluster (@arc ref). Using our selected papers from the training dataset, we downloaded the entirety of each paper's source HTML using the command line tool wget. This allowed us to use the source HTML multiple times for updated analyses without the need to re-query the ASM webservers numerous times. Next, we performed cleaning of the HTML using R packages rvest (@rvest ref) and xml2 (@xml2 ref) to get the desired portions of the paper from the HTML including the abstract, the body of paper, all tables and figureswith captions, as well as the side panels for all papers, but especially those containing the data availability statements in papers published after the 2023 change in webpage format (see above). Then we removed unnecessary text using R packages tm(text manipulation)(@tm ref) and textstem (@textstem ref), as well as converting all text to lowercase, and the removal of digits and non-alphabetic characters such as whitespace. To have the fewest number of unique words, we lemmatized (sort words by grouping inflected or variant forms of the same word) words to trace them back to their root words and eliminate any possible issues with word tense. After this, we created and counted our 'tokens', phrases of up to 1-3

consecutive words from the text of the paper using R package tokeinziers (@tokeinziers ref). Towards the goal of the fewest meaningful number of words, we used the 'Snowball' (@snowball ref) dictionary of 'stop words' to remove non-meaningful words such as articles 'a', 'an', and 'the'. We removed the 'space' character with an underscore in multi-word tokens for ease of procesing, and created a count table for the tokens in each paper.

Once the tokens in each paper were counted, we transformed the data into a sparse matrix format useable by the R package mikropml (@mikropml ref), using R packages caret and dplyr (@caret ref, @dplyr ref). Tokens were filtered to those which appear in greater than one paper. This allows comparison between papers by the model. We removed near zero variants (tokens with frequency very close to zero) as well as collapsing perfectly correlated tokens (tokens that always appear together) using R packages caret and mikropml to reduce model complexity. The data was then simplified to keep only the following variables; tokens, frequency, journal information, and hand identified NSD and DA variables. This simiplified sparse matrix data had the mean and standard deviation calculcated and saved for the frequency of each token to later apply a z-scoring method to future data to be predicted by the model.

Training of the DA and NSD Models We trained two random forest machine learning models using mikropml's "run_ml" function, one to determine if a paper contained new sequence data (NSD), and another to determine if the paper had data availabile (DA). The mikropml "run_ml" function uses methodology described by Topcuoglu et al (@topcuoglu2020) to split data for model training. Random forest models have one hyperparameter to tune, the mtry value. We began with mtry values of 100, 200, 300, 400, 500, and 600, to find peak hyperparameter performance given *N tokens*. We trained the models multiple times in accordance with existing methologies, first to find the optimal Area Under the Receiver-Operator Curve (AUROC) value for each model with N=100 seeds. Then to find the best mtry performance for each model, with N=1 seed. Finally, with N=1 seed to train each final model for use on experimental data.

Preparation of the Experimental Dataset To fully answer our research questions, we created a larger database with N = 155779 papers curated from reference datasets Crossref, NCBI, Scopus, and the Web of Science (@crossref, @ncbi, @scopus, @wos). These papers span all twelve ASM journals of interest from the start of 2000 to the end of 2024. Once database was curated, we applied the same steps to ready papers for application of machine learning models as the model training datasets. See above for descriptions of webscraping html, cleaning html, removing unnecessary text, and creation of token count table for application in each of the machine learning models to determine the NSD and DA statuses for each paper. Once the frequency count tables were prepared for each paper, a z-score was applied using the saved data from each model appropriately, using the formula ((observed_token_frequency - model_token_frequency_mean)/(model_token_frequency_sd)). This z-scoring formula was applied to standardize the frequency of each token. Only tokens included in the machine learning models were retained in

experimental datasets. Finally, each model was deployed on each paper to determine its NSD and DA status.

  - Statistical methodology
    * negative binomial modeling using r package MASS
      · fixed modeling
      · log transformation of age.in.months
      · 95% CI

- Supplemental Material file list (where applicable)
- Acknowledgments
- References
- Figures/Tables/stats to make/get
  - Table showing # of papers from each journal in training dataset
  - Table showing # of papers from each year in training dataset
  - % of papers nsd yes total (training)
  - % of papers da yes total (training)
  - table of conditions to add to the methods of classification?