A Machine Learning-Enabled Open Biodata Resource Inventory from the Scientific Literature

Heidi J. Imker[1, 2] (ORCiD: 0000-0003-4748-7453), Kenneth E. Schackart III[1, 3] (ORCiD: 0000-0002-1658-3699), Ana-Maria Istrate[4] (ORCiD: 0000-0002-7953-5168), Charles E. Cook[1] (ORCiD: 0000-0002-4145-8048)

[1] Global Biodata Coalition, 12 quai Saint-Jean 67080, Strasbourg, France

[2] University Library, University of Illinois at Urbana-Champaign, Urbana, Illinois 31821, USA

[3] Department of Biosystems Engineering, The University of Arizona, Tucson, Arizona 85721, USA

[4] Chan Zuckerberg Initiative, Redwood City, California 94063, USA

Abstract

Modern biological research depends on data resources. These resources archive difficult-to-reproduce data and provide added-value aggregation, curation, and analyses. Collectively, they constitute a global infrastructure of biodata resources. While the organic proliferation of biodata resources has enabled incredible and novel research, sustained support for the individual resources that make up this distributed infrastructure is a challenge. The Global Biodata Coalition was established by research funders in part to aid in developing sustainable funding strategies for biodata resources. An important component of this work is understanding the scope of the resource infrastructure; how many biodata resources there are, where they are, and how they are supported. To complement existing registries, which require self-registration and/or extensive curation, we sought to develop a methodology for assembling a global inventory of biodata resources that could be periodically updated with minimal human intervention. The approach we developed identifies biodata resources using open data from the scientific literature. Specifically, we used a machine learning-enabled natural language processing approach to identify biodata resources from titles and abstracts of life sciences publications contained in Europe PMC. Pretrained BERT models were fine-tuned to classify publications as describing a biodata resource or not and to predict the resource name using named entity recognition. To improve the quality of the resulting inventory, low-confidence predictions and potential duplicates were manually reviewed. Further information about the resources were then obtained using article metadata, such as funder and geolocation information. These efforts yielded an inventory of 3112 unique biodata resources based on articles published from 2011-2021. The code was developed to facilitate reuse and includes automated pipelines. All products of this effort are released under permissive licensing, including the biodata resource inventory itself (CC0) and all associated code (BSD/MIT).

Introduction

Scientists have long undertaken major infrastructure projects that examine large-scale scientific questions. Such initiatives often require sustained long-term support, frequently from a defined set of funders, for decades. The physical sciences have been especially adept at establishing infrastructures to

produce data that are critically important for furthering scientific understanding. Typically, these projects, for example the large hadron collider at CERN and the James Webb Space telescope, are tangible structures or instruments that have well-defined physical locations (whether on earth or in space). Because they are tangible objects, the funders and the taxpayers who ultimately support the infrastructure can readily understand both how funds are spent and the necessity of long-term support in order to ensure that returns on the high initial investments are maximized.

In contrast, the life sciences' data infrastructure is highly distributed and largely virtual. There are thousands of distinct systems that provide access to structured biological data, collectively referred to as biodata resources. These resources archive difficult-to-reproduce data and also provide added-value aggregation, curation, and analysis to those archived data. Biodata resources are found throughout the world, vary in scale, and are supported by hundreds of funding bodies, institutions, and charitable foundations. This distributed infrastructure has grown dramatically over the past three or four decades as technological advances, such as in nucleotide sequencing, enabled exponential increases in the amount and types of data generated. However, and again unlike physical sciences, growth has been organic and driven locally as individual researchers and institutions sought or provided funding to create each data resource, with new resources joining the infrastructure individually when they begin exchanging data with other resources.

The impact of biodata resources on life science research has been immense, and a number of efforts have been launched in recent years to improve their coordination and long-term sustainability. In Europe, ELIXIR was established in 2013 as an intergovernmental organization to coordinate life science data infrastructure. As part of ELIXIR's mission, they identified a set of European Core Biodata Resources that are of "fundamental importance" to research and show "wide applicability and usage" based on a set of quantitative and qualitative indicators. Initially, 19 resources were identified (now 22 after additional selection rounds), and literature mentions and citations show incredible reach. In an analysis of Europe PMC's full text articles, 17% were found to refer to a core resource [1].

Despite their critical importance for the life sciences research endeavor, biodata resources are usually funded precariously through short-term grants (generally 3–5 years) [1–3]. Research funders provide support for many of the biodata resources that comprise the biodata infrastructure, and they recognize both the need for long-term support of data resources and the challenges associated with creating long-term funding streams for such support [4]. While not all resources should live on in perpetuity, there is collective interest in establishing alternate funding mechanisms to stabilize the resources that make up the infrastructure [5]. In recognition of this challenge, research funders supported creation of the Global Biodata Coalition (globalbiodata.org) to aid them in coordinating funding for biodata resources and to develop mechanisms to more efficiently fund the biodata infrastructure. A basic requirement for coordinating support for this infrastructure is to understand its scope: how many biodata resources are there and where are they located? However, because biodata resources have been developed and managed independently of each other, this global overview is missing.

There have been many efforts to catalog biodata resources over the years. An early example is DBCat, launched in 1999 by the EMBnet branch Infobiogen, which used a combination of general web searches, journal review, and contributions from resource providers to assemble a list of 511 resources [6]. Another effort in the biodiversity community found over 600 biodiversity information projects

between 2005-2006 based on 100 hours of consulting effort [7]. There are also partial lists created by research funders [8], academic libraries [9], scholarly publishers [10], and Wikipedia [11]. The journal *Nucleic Acids Research* also maintains a catalogue of primarily molecular biology-related databases, the vast majority of which are described in one of the annual database issues of that journal [12].

Blair et al. noted the challenge of maintaining catalogs, which become quickly outdated if static, but likewise are subject to sustainability challenges themselves when dedicated staffing and funding are required. Indeed, the largest collection of molecular biology databases is Database Commons, hosted by National Genomics Data Center in Beijing China, which has developed its 5000+ record collection with contributions from over 50 curators. An alternative is encouraging resource owners to register their own resources. Currently, there are several options for such registration, including re3data.org [13], FAIRsharing [14], and the SciCrunch Registry [15]. All three actively encourage registration and include between ~ 1500 to ~ 3000 biodata resources, depending on interpretation of categories.

Given the interest in biodata resources from many different perspectives and the challenge of being able to document the ever-growing global life sciences infrastructure, we sought a method of assembling a global inventory of biodata resources and creating a process that would enable this inventory to be periodically updated. Here we describe the results of a reproducible, machine learning-enabled method to create this inventory by identifying biodata resources described in scientific articles between 2011-2021, thereby providing an open and updateable basis for describing the life sciences' highly distributed data infrastructure. The inventory developed, which contains 3112 resources, represents a use case-focused practical application of machine learning to address a question of interest to research funders and other stakeholders who support and use biodata resources across the globe. To facilitate reuse, we have released the inventory under CC0 licensing along with code available under BSD/MIT licensing; the code includes the machine learning steps in an automated pipeline plus scripts that extract value-add information about the identified resources from the metadata of associated articles. Along with a presentation of the methodology, we also provide a preliminary analysis of the inventory to demonstrate the potential for its reuse and augmentation.

Methods

Design Overview

In order to create an open, reproducible inventory we needed a large source of open data that contains information about biodata resources and is structured enough to enable programmatic access. Text mining the scientific literature has been used to locate resources in previous studies, and Wren et al. combined this strategy with crowdsourcing to classify over 20,000 URLs extracted from MEDLINE abstracts, including 4757 that were designated as databases [16]. While Wren et al. did not identify resource names or locations, the high number of results suggested that scientific abstracts are a viable data source for identifying a large cache of resources. Furthermore, articles in centralized literature services are associated with high-quality metadata available via robust APIs. This means for articles that specifically describe a biodata resource, associated metadata such as title, abstract, authors, author affiliations, funders, and citations, can be used to extract or infer information about the biodata resource itself, even when full text articles are paywalled. Additional metadata associated with the resource URL

extracted from the abstract (e.g., HTTP status and IP location) also allows collection of a set of useful characteristics for each inventoried resource. While there are limitations to this strategy since not all resource owners publish articles describing their biodata resources, many do, and we hypothesized that we could not only create a large inventory that could be of use in and of itself, but freely releasing the inventory and the associated code would allow for reproducibility and extension by others interested in subsetting or augmenting the inventory for other purposes.

Europe PMC is a large data resource of life sciences literature with an API allowing full access to the entire resource [17]. Using the strategy described above, we developed a targeted query to retrieve from Europe PMC a corpus of articles and then tested and used a machine-learning based approach to identify biodata resources named in this corpus (Figure 1). Openly available pretrained language models are currently state-of-the-art in natural language processing (NLP), achieving high performance on a variety of tasks such as Named Entity Recognition (NER), Question Answering, summarization, and machine translation. These models have also been adapted to the biomedical field by pretraining on domain-specific corpora (e.g., BioBERT, SciBERT, PubmedBERT). We defined two tasks: article classification, which aimed to classify a research article (based on the title and abstract) as being about a biodata resource or not, and NER, which extracted the exact mentions of biodata resources from text. We experimented with several of these pretrained language models relevant to the biomedical field by fine tuning them on these two tasks. We used a regular expression algorithm to extract URLs corresponding to the biodata resources and checked their HTTP response statuses and locations. With articles and their biodata resources defined, we then accessed additional metadata to further characterize individual resources, as mentioned above. This entire workflow was automated and made reproducible by implementing it in a Snakemake pipeline [18]. To maximize utility across a wide variety of potential users, the results have been made available as shown in the Open Science Products Table (Table S7).

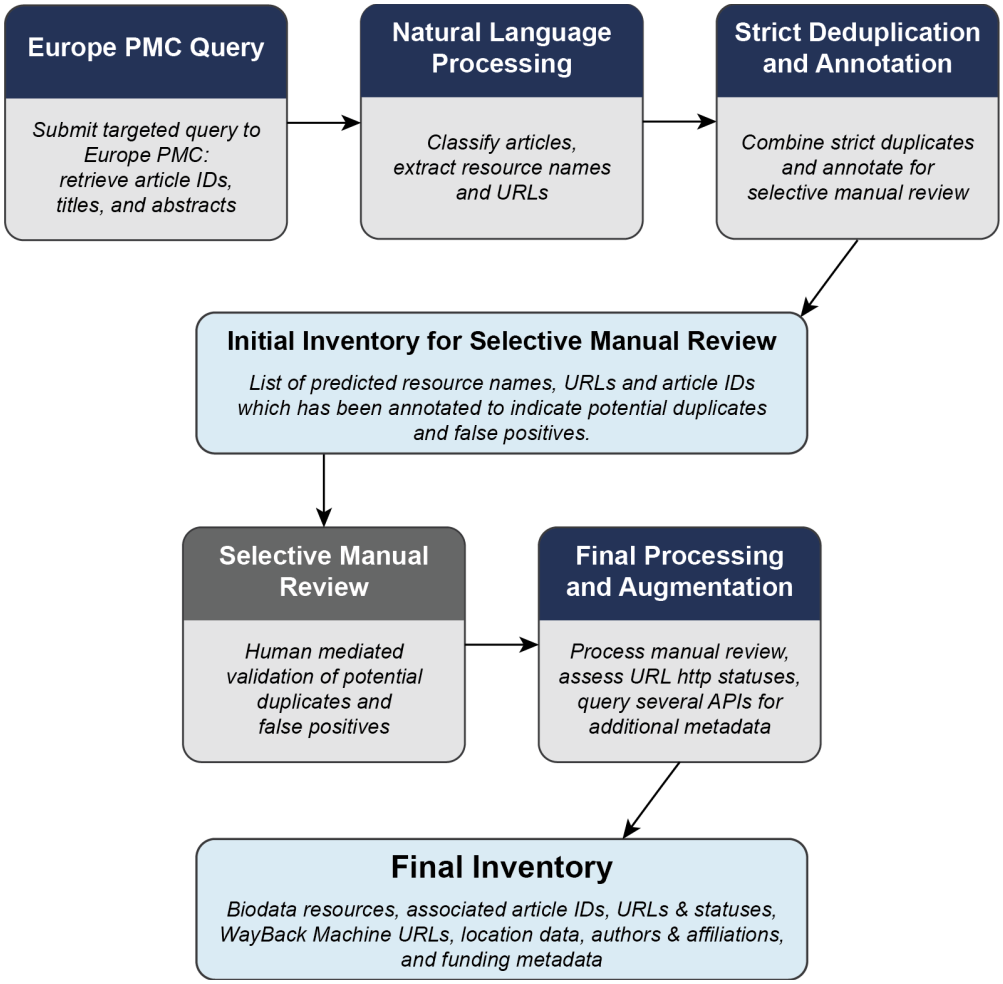

Figure 1. Study design for creating an inventory of biodata resources starting with a targeted query of articles found in Europe PMC, predicting biodata resources from article titles and abstracts using NLP, then reviewing and augmenting with additional metadata from Europe PMC to create the final inventory.

Open Science Implementation Plan

At the start of the project, an Open Science Implementation Plan was created in order to clearly articulate our Open Science goals. The plan guided decision-making, allocation of time and resources, and the ultimate products of the project. The plan has four components: Reproducibility Standards, Code Standards, Data Standards, and External Review/Validation [19]. An account of our efforts to adhere to this plan, including a detailed description of steps taken to follow reproducibility standards outlined by Heil et al., are described in a companion article [20,21].

Working Definitions

Creating the inventory required definitions of “biodata” and “biodata resource” explicit enough to allow human curators to evaluate articles and create a training dataset for the machine learning models. For this inventory, which is intended to be of use to funders of basic research worldwide, we specifically excluded clinical and patient data resources that appeared to be aimed at clinical or

diagnostic use rather than for use in research. We reviewed both formal ontologies and generic definitions related to the life sciences and the basic sciences (Table S1) to assemble a working definition of *life sciences biodata* that reflected the objective of the inventory:

Working definition of "life sciences biodata"
Biodata are created through studies of living organisms and their associated life processes through research conducted for the specific purpose of acquiring fundamental knowledge; this knowledge forms the basis of testable theories that aim to increase our ability to understand, interpret, and predict the phenomena that impact these organisms and processes.

To define "biodata resource" we likewise reviewed existing formal and general definitions (see Table S2), including those from re3data.org, ELIXIR, the US National Institutes of Health, and the US Department of Energy as well as standards such as the Biomedical Resource Ontology and W3C's Data Catalog Vocabulary (DCAT):

Working definition of a "data resource"
An online source of structured data. The data cannot be a copy readily obtained from another resource (e.g., must be primary data or data annotated, curated, or otherwise augmented with value-added elements that are unique to the resource). The resource must have a distinct name and interface for browsing, searching, querying, viewing, and/or downloading the data within. Mechanisms such as an API may be the main access method, but there still must be a distinct online presence that provides information about the data available. Analysis tools may be provided, but information about and access to the underlying, unique data must be clearly available to users visiting the resource.

A particular challenge for this project was distinguishing biodata resources, which explicitly provide access to data, from tools that provide analysis or visualization of data, either through an inaccessible background database or via input by a user. In cases where the resource appeared to be both a source of biodata and a tool, it was included in the inventory.

Data Sources

To ensure that all project data are open and freely distributable, and that the inventory may be updated programmatically in future, only open public data accessible via an API was used. The Europe PMC API was accessed using the Python requests library (v2.27.1)[22] to gather articles that potentially describe a biodata data resource. The query string was developed to locate abstracts that contained a URL and a small set of focused keywords that suggested the articles described a data resource while excluding retracted articles and those that include general URLs (e.g. for clinical trial registrations). The final query string used is shown in Figure S1 and provided in the project's data deposits and git repositories.

Data retrieved from the re3data.org and FAIRsharing APIs are licensed for reuse and were used to benchmark the resulting inventory. Finally, Wayback Machine URLs and geo coordinates were retrieved via the Internet Archive Wayback Availability, ipinfo, and ip-api APIs as indicated in Table S3.

Natural Language Processing (NLP) Tasks

To generate the inventory from open data several natural language processing (NLP) methods were employed. ML models were trained to perform article classification and named entity recognition (NER) to identify articles describing biodata resources and extract their names. A regular expression is used to extract the URLs from the abstracts of predicted biodata resources. This workflow for the NLP tasks is shown in Figure 2, and details about the ML methods are covered in the section below.

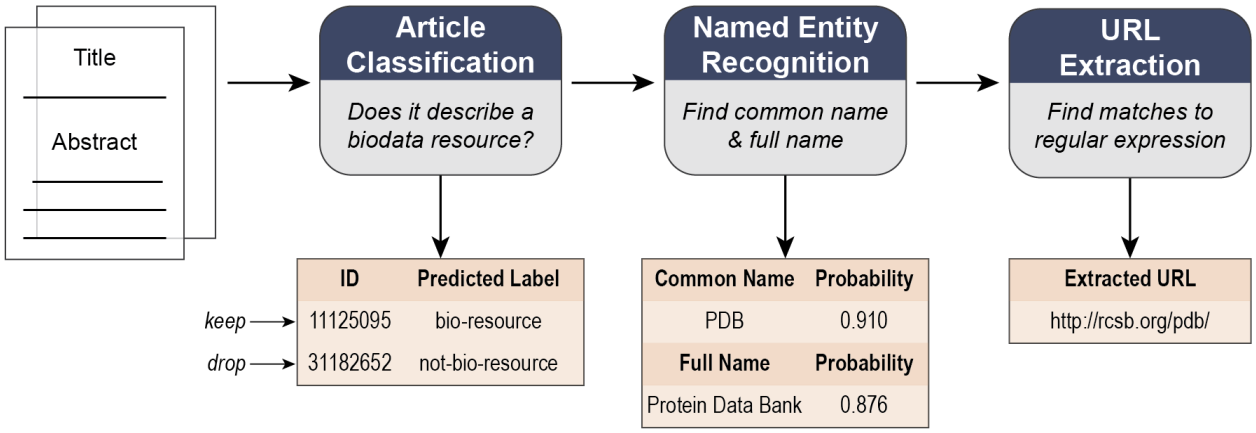

Figure 2. Workflow for the process of extracting mentions and URLs of biodata resources from a research article. The article title and abstract are first fed into an article classification model, which determines if the article is about a biodata resource or not. Articles that are classified as being about a biodata resource are then fed into an additional Named Entity Recognition (NER) model which extracts the name(s) of the biodata resource and reports the average probability score of the tokens constituting the predicted name. A regular-expression based URL extraction algorithm separately extracts URLs from the abstract. In this example, the NER model extracted both a "full name" (Protein Data Bank) and a "common name" (PDB) for the resource, which are both valid in this case.

Training Data

To create the training dataset needed to develop a classifier via machine learning, a random sample of records returned from the query above was selected for manual review. The training set was created in two phases, with the first containing 638 records and second containing an additional 996 records for a total of 1634 records in the training dataset. Article titles and abstracts were independently reviewed by two curators in each phase and classified as either describing a biodata resource or not describing a biodata resource based on the developed definitions (see above). Initial inter-annotator agreement was high (89.4%) and increased to 97.1% once conflicting scores were reviewed, discussed, and reclassified. Challenges fell into two main issues where it was difficult to distinguish between biodata resources and other resources that 1) are available solely as a tool or 2) belonged to a different disciplinary area (such as clinical health records).

We kept the entries where both curators agreed on the article classification label (either positive or negative, n = 1587) and used this as a training dataset for the article classification task. For articles manually classified as describing a biodata resource, mentions of biodata resources in the title and abstract were identified, including “common names” (e.g., PDB) and “full names” (e.g., Protein Data Bank). This curated set of mentions was used for the NER task. Both training datasets were split into 70% training, 15% validation, 15% test (hold-out) for article classification (Table 1) and NER tasks (Table 2).

Table 1. Training dataset splits for the article classification task

	Train	**Validation**	**Test**	**Total**
Positive labels	337 (30.4%)	61 (25.6%)	80 (33.5%)	478
Negative labels	773 (69.6%)	177 (74.4%)	159 (66.5%)	1109
Articles	1110	238	239	1587

Table 2. Training dataset splits for the Named Entity Recognition (NER) task

	Train	**Validation**	**Test**	**Total**
Articles	306	66	66	438
Biodata resource mentions	1192	269	293	1754

Models

Two machine learning models were fine-tuned to automate the process of 1) classifying research articles and 2) extracting mentions of biodata resources from those predicted to describe a biodata resource. Given a paper’s title and abstract, the article classification model classifies the paper as being about a biodata resource or not. If an article receives a positive score, it is then passed through the NER model, which extracts the common name and full name of the biodata resource from the text, if they are present. A confidence score, computed as the average probability among the tokens constituting the mention, is also output (Fig. S2). BERT (Bidirectional Encoder Representations from Transformers) performs well on a variety of NLP tasks and several BERT derivatives have been pre-trained on biomedical corpora, making them excellent candidates for this project. For both the article classification and NER task, BERT itself and 14 other BERT model variations available on Hugging Face Hub (https://huggingface.co/) were fine-tuned and evaluated to select the highest performing model (Table 3 and citations therein).

Table 3: Pre-trained models that were fine-tuned for the article classification and NER tasks

Model	**Hugging Face Model Name**	**Citation**
BERT	“bert-base-uncased”	[23]

BioBERT	“dmis-lab/biobert-v1.1”	[24]
BioELECTRA	“kamalkraj/bioelectra-base-discriminator-pubmed”	[25]
BioELECTRA-PMC	“kamalkraj/bioelectra-base-discriminator-pubmed-pmc”	[25]
BioMed-RoBERTa	“allenai/biomed_roberta_base”	[26]
BioMed-RoBERTa-CP	“allenai/dsp_roberta_base_dapt_biomed_tapt_chemprot_4169”	[26]
BioMed-RoBERTa-RCT	“allenai/dsp_roberta_base_dapt_biomed_tapt_rct_500”	[26]
BlueBERT	“bionlp/bluebert_pubmed_uncased_L-12_H-768_A-12”	[27]
BlueBERT-MIMIC-III	“bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12”	[27]
ELECTRAMed	“giacomomiolo/electramed_base_scivocab_1M”	[28]
PubMedBERT	“microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract”	[29]
PubMedBERT-Full	“microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext”	[29]
SapBERT	“cambridgeltl/SapBERT-from-PubMedBERT-fulltext”	[30]
SapBERT-Mean	“cambridgeltl/SapBERT-from-PubMedBERT-fulltext-mean-token”	[30]
SciBERT	“allenai/scibert_scivocab_uncased”	[31]

Article Classification Task

The pre-trained models were fine-tuned on the article classification task. The model with the highest performance on the validation set was selected and used to generate the inventory. In order to consider both title and abstract for classification the title and abstract were concatenated with a space character between fields to create a contiguous string. XML tags were removed using regular expressions, while adding white space after punctuation if not present after tag removal. The resulting input string was tokenized using a pre-trained tokenizer associated with the specific pre-trained model to be used for classification. Tokenized input was then passed through the pretrained model module to obtain context embeddings, which were subsequently fed into a linear classification layer that performs binary classification. This process classifies an article, based on title and abstract, as describing a biodata resource or not (Figure 3A).

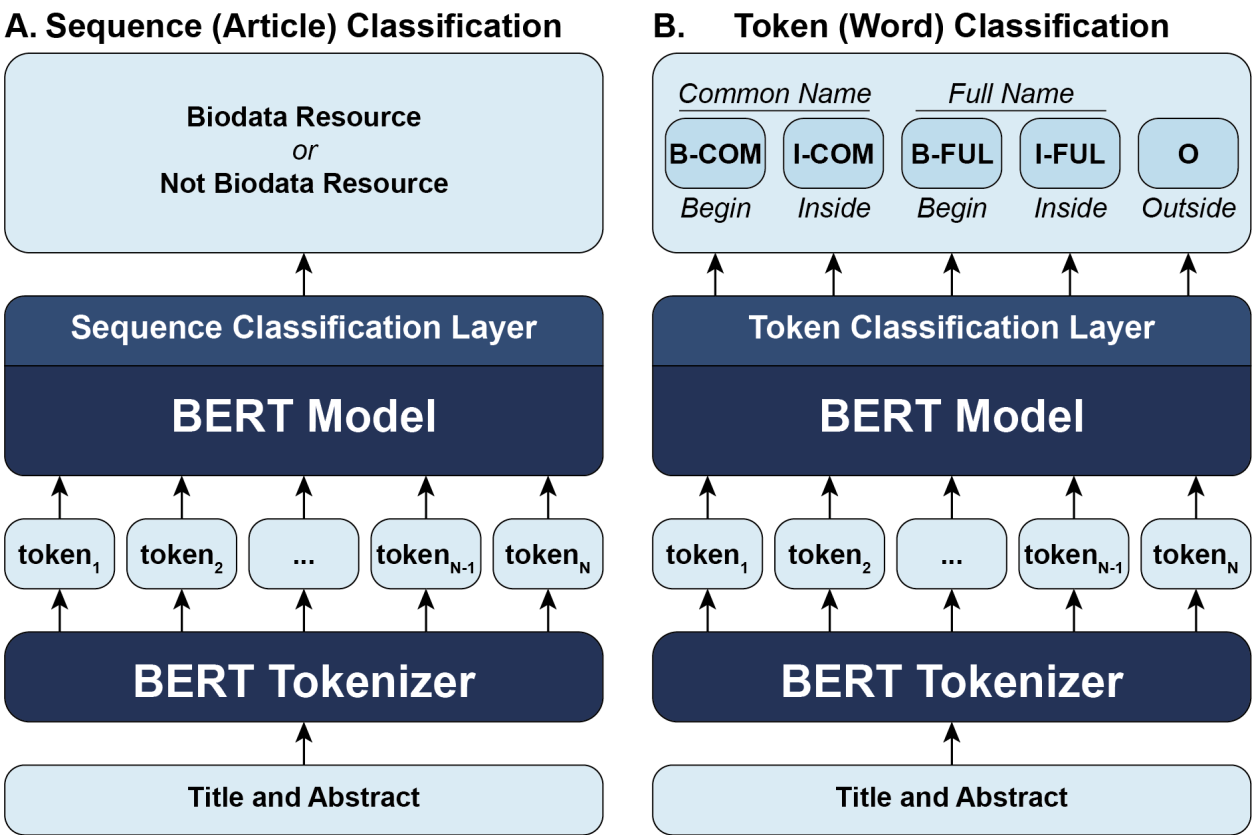

Figure 3. Model architecture for A) sequence classification used to classify articles based on title and abstract and B) token classification used to perform NER to obtain the resource names. Each architecture shows the possible classification labels resulting from model prediction. Additional details for token classification are shown in Figure S2.

Classification performance was evaluated on the validation set using precision (eq. 1), recall (eq. 2), and *F*1 score (eq. 3). For calculating performance metrics, an article that describes a biodata resource that is correctly classified is a true positive (TP), and if incorrectly classified is a false negative (FN). An article that does not describe a biodata resource that is correctly classified is a true negative (TN), and if incorrectly classified is a false positive (FP). Precision is the proportion of those articles predicted to describe a biodata resource that are indeed describing a biodata resource. Recall is the proportion of articles that describe a biodata resource that are correctly classified. *F*1 score is the harmonic mean of precision and recall.

$$precision = \frac{TP}{TP+FP} \quad (1)$$

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = \frac{2*precision*recall}{precision+recall} \quad (3)$$

Models were trained for a maximum of 10 epochs. For each model the model checkpoint with the highest precision (regardless of epoch number) was saved. Precision was chosen over *F*1 and recall for the classification task to reduce the number of false positives. The precision scores of all the fine-tuned models were compared to select the most performant model, which was then used for classification of unlabeled data.

Named Entity Recognition (NER) Task

The same pre-trained models evaluated for the article classification task were fine-tuned for the NER task (Table 3). To accomplish this, the linear sequence classification layer was replaced by a token classification layer (Figure 3B and Figure S2). Input (training and validation) data was tagged using the following BIO scheme [32]:

- B-COM: token is the start of a sequence corresponding to a common name
- I-COM: token is a non-start part of a sequence corresponding to a common name
- B-FUL: token is the start of a sequence corresponding to a full name
- I-FUL: token is a non-start part of a sequence corresponding to a full name
- O: otherwise

NER performance was evaluated using partial-match entity level metrics (precision, recall, *F*1 score) on the validation set. As with the classification task, models were trained for a maximum of 10 epochs. For each model architecture the trained model checkpoint with the highest *F*1 score was saved. The model with the highest *F*1 score on the validation set was then used for downstream tasks.

Model Implementation Details

The Huggingface framework was utilized to load the BERT model architectures [33]. The Huggingface's AutoModelForSequenceClassification module was used to fine-tune the BERT models on the article classification task, and the AutoModelForTokenClassification module to fine-tune the models on the NER task. The hyper-parameters used during fine-tuning of the models can be found in Table S4. We used the Adam optimizer for training and the seqeval module for computing partial match entity-level metrics. Models were trained for a maximum of 10 epochs. Full model implementation details are available in the Supplemental Materials. All machine learning code was implemented in Python v3.8.12 and used the following third-party packages: Datasets v1.18.3, Natural Language Toolkit (NLTK) v3.6.1, NumPy v1.19.2, Pandas v1.2.4, pytest v6.2.4, scikit-learn v0.24.1, PyTorch v1.9.0, tqdm v4.63.0, and transformers v4.16.2 [33–41].

Mid-Project Evaluation and Iteration

To assess how well the predictions were performing in practice and if any improvements could be made to any part of the process before we continued we conducted a manual evaluation midway through the project. At that point, SapBERT resulted in the highest best *F*1 score of the 15 models tested for the classification task (Table S5). Using the results of that model, a curator assessed precision on a 10% random sample (n = 468) of the predictions to assess the correctness of the article classifications and NER extracted terms. Classification was determined to be either correct or incorrect. For the NER outputs, the evaluation determined if the extracted common and full names were correct, partially

correct, or incorrect. This step helped us confirm that machine learning was indeed viable for this project and also helped us determine that a selective manual evaluation of low-scoring predictions would still enhance the overall quality of the inventory. To determine which predictions qualify as “low-scoring” for this first inventory and updates in the future, we used the results of the midway evaluation to determine a threshold probability of < 0.978 as “low-scoring,” where 0.978 was the average probability for names determined by a curator to have been correctly predicted in the 10% random sample (see additional details in Manual Evaluations in the Results section). Therefore, any resource whose highest scoring predicted name (“best name,” regardless if common or full) has a probability below this threshold is annotated to be manually reviewed by a curator before the inventory is finalized. In addition to the threshold, this step also helped us determine what iterative improvements could be made to refine the overall strategy, such as revising the input query (e.g., excluding “onlinelibrary.wiley.com” URLs) and adding conditionals to the models (e.g., predicted name length > 1).

Post Processing

Best Name Determination

As detailed above, for each article the NER model attempts to predict both full and common names and may output multiple predictions, each of which is associated with a probability score, where the higher the score the greater the confidence in the predicted name. To determine the highest quality names, the probability scores for named entities of each type were compared to determine the “best common name” and “best full name”. These probability scores of these two named entities were compared to choose the best overall name “best name” (that with the highest probability score).

Automated Deduplication

Many resources publish repeatedly, for instance, to provide updates about the resource. Consequently, the raw inventory contained duplicate records from several articles describing the same resource. A first step toward deduplicating the inventory was performed by identifying records that had the same predicted best name and same extracted URL (ignoring differences due to trailing slashes or “http:” vs “https:”). These duplicate resources were merged with the PMIDs of each original article and retained along with the title-abstract text and publication date of the most recently published article.

Annotation for Selective Manual Evaluation

Up to this point all steps were automated. However, based on the results of the mid-project evaluation, we realized that it would be advantageous to conduct a selective manual evaluation of some predicted resources to improve the overall quality of the inventory. In preparation for this step, a script within the pipeline added a new column with the variable “low_prob” for any resource whose best name probability < 0.978, the value determined in the mid-project evaluation as the average of correctly predicted names in the 10% random sample. This served as a flag to aid the curator conducting the manual evaluation. Additionally, while the automated deduplication was able to merge any articles with exact names and exact URLs matches, we were aware of suspected duplicates (e.g. variable names such ''FANTOM'' and ''FANTOM5'' sharing the same extracted URL while variable URLs such as

http://appris-tools.org and http://appris.bioinfo.cnio.es share the same predicted name ''APPRIS''). Deduplication on either name or URL (as opposed to both) would have led to erroneous mergers. For example, ''Seed'' and ''SEED'' are two different resources, while there are at least three distinct ''PED'' resources and two distinct resources for ''SMART.'' Instead, to account for potential duplicates, columns were also generated for matching best names (flagged with “duplicate_names”) or matching extracted URLs (flagged with “duplicate_urls”) for evaluation by the curator. While more complex, automated procedures may be warranted in the future as the inventory grows over time, these cases were relatively few and we judged this strategy to be sufficient for the time being.

Selective Manual Evaluation Procedure

With steps implemented above, a preliminary inventory with records flagged for manual review was generated as a CSV file. A curator then reviewed each flag to determine if low probability records should be removed from the inventory and if potential duplicate records should be merged within the inventory. This review was done in Microsoft Excel, with data validation applied to a set of predetermined outcomes (e.g. ''remove'', ''merge'', etc.). Importantly, no corrections to the predicted names or URLs were made; thus all values within the inventory are the output of the machine learning pipeline, which reduces the confusion that could result if the inventory were a mixture of ML-generated and human-generated names/URLs (especially in future updates). Review guidelines were developed to help standardize handling of edge cases (see Table S7. Open Science Products). The resulting manually reviewed file was added into the directory for subsequent processing. The pipeline first ensures all flagged records contain only valid review values and then removes or merges the appropriate records before moving on to metadata augmentation.

Metadata Augmentation

Once we had the final prediction script and the results of the manual review processed, the Europe PMC API was once again queried using PMIDs to retrieve author affiliations, author names, grant IDs, grant agency name, and citation counts (i.e. via metadata elements 'affiliation’, 'fullName', 'grantID’, 'agency’, 'citedByCount', respectively) for each article associated with the biodata resources.

URL Processing

Biodata resources may be impermanent for reasons that include loss of funding, loss of key personnel, and technological change leading to deprecation, and we were interested in establishing if the URL provided in the abstract was still viable. Accordingly, the extracted URLs were checked for viability through standard HTTP status calls using the Python requests and urllib3 (v1.26.8) [42] libraries and the returns (e.g. 200 OK, 404 Not Found, etc.) were recorded in the inventory. Extracted URLs were tested three times, with each attempt allowing 5 seconds for a response. The second attempt is submitted immediately after the first, while the third attempt is submitted after a one second delay.

Web archives offer a chance to locate snapshots of previously available websites [43]. To mitigate current and anticipate future availability issues, we used the Internet Archive’s Wayback Machine API to check URLs for the presence of archived sites. While the Wayback crawler is often unable to access the data itself, these snapshots provide views of HTML pages such as the home page, search

interface, etc. which provide important context in the absence of the live site. For the biodata resources in this inventory the most recent Wayback Machine URL was recorded for successful returns; for live URLs not represented in the Wayback Machine the URLs were submitted for archiving and the associated Wayback URL recorded. While the Wayback Machine is able to crawl and archive the majority of sites represented here, sites behind firewalls or those that prohibit crawlers cannot be archived.

Resource Geolocation

We use two methods to identify the location of the biodata resource. The first is the location as determined from the IP address, which suggests a physical location for the infrastructure. For those URLs that return a status less than 400, the IP address is obtained from the host name. In an attempt to geolocate the IP addresses, ipinfo and ip-api are queried to request the IP address country and coordinates. Two APIs are used since neither is complete, and querying multiple APIs increases the chance of successful geolocation. When a location is successfully obtained from either API, the country and coordinates are recorded.

Because of the global nature of the life sciences research enterprise, physical location alone may not reflect collaboratively developed resources. Therefore, we also extracted country names following ISO 3166 from the author affiliations available from the Europe PMC metadata.

Workflow Management

The Snakemake workflow manager is used to automate the process described above in two pipelines. The first pipeline performs data splitting, model training (including model selection and evaluation), prediction, and downstream processing prior to selective manual review. After the selective manual review this pipeline resumes for final processing. This process, excluding model training and selection, is shown in Fig. 1. A second pipeline was also developed to facilitate updating the inventory in future using new queries to Europe PMC and the previously fine-tuned models.

Analysis of the Final Inventory

With the final inventory established and additional metadata elements added, we carried out analyses on location, funders, and text-mining related metadata to provide a preliminary analysis of the resources identified and explore future opportunities to reuse the inventory. To evaluate funders, curation was done to determine the country of origin for agencies mentioned >2 times (see Supplemental Methods). Additionally, to get a sense of how well this method complements other methods, we compared the biodata resources in this inventory with life science data resources retrieved via the APIs of two prominent registries, re3data.org [13] and FAIRsharing [14]. All analyses were performed in R using the packages argparse, dplyr, europepmc, forcats, ggmap, ggplot2, glue, gt, httr, jsonlite, magrittr, maps, purrr, RColorBrewer, readr, scales, stringr, tibble, tidyr, and xml2 [44–63]. The data resulting from these analyses, as well as the scripts to perform them, are available in GitHub and archived in Zenodo along with the code and data for developing the inventory.

Results

Overview

We initially retrieved 21716 articles from Europe PMC using all data sources, but later restricted this to the 20880 which had PMIDs available since we found that the metadata associated with articles lacking PMIDs was often insufficient for further analyses. During the article classification task the model predicted a negative label for 16588 articles while 4292 articles were predicted to describe a biodata resource. Of those with a positive label, the NER model extracted at least one name (common or full) for 4006 articles. Those without at least one complete URL were removed, resulting in 3940 articles. After deduplicating articles that had the same "best name" and same extracted URL, 3565 potential resources were identified in the preliminary inventory. After selective manual review to evaluate and remove erroneous low probability resources and merge additional duplicates, the final inventory contains 3112 biodata resources.

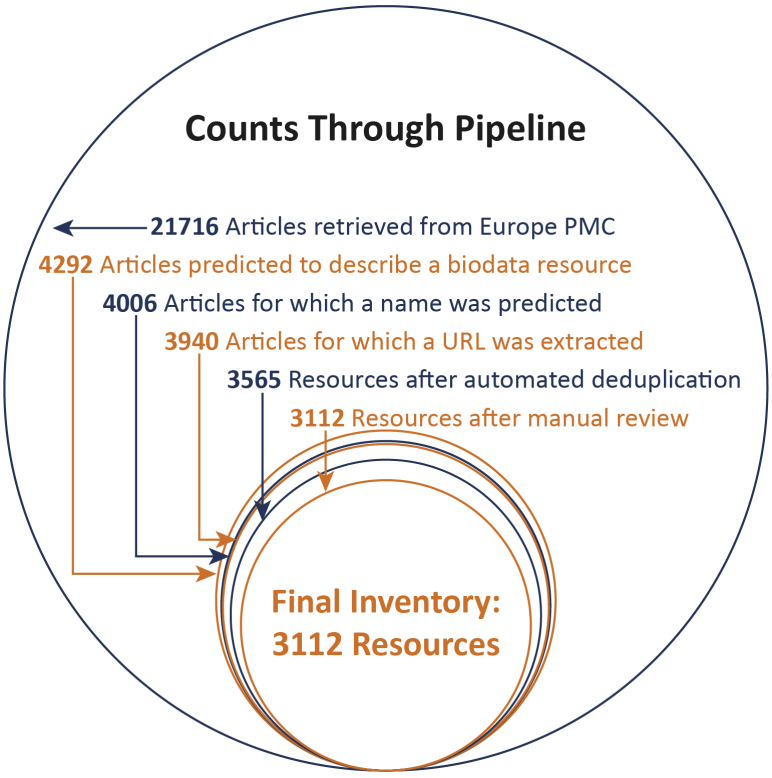

Figure 4. Statistics on article and resource counts as they pass through the pipeline. The area of each circle is proportional to the count at that step. Most attrition occurs during the initial classification step that predicts which articles describe a biodata resource, while losses are relatively low as additional steps build and finalize the inventory.

NLP Tasks Metrics

For each of the article classification and NER tasks we divided the training datasets into 70/15/15 splits used for training, validation, and testing, respectively. Models were evaluated using Precision, Recall and the *F*1 scores. All pretrained models were fine-tuned and evaluated using the same training

and validation sets. Models with the highest performance when run using the validation data were chosen for final implementation. Precision was used for comparing the article classification models, and *F*1 score was used for comparing the NER models. Performance of all fine-tuned models on both validation and test sets are provided in Tables S5 and S6. For the article classification task BioMed-RoBERTa-RCT had highest precision on the validation set and had a precision of 0.975, *F*1 score of 0.821, and recall of 0.719 on test data that had not been seen during training. For the NER task, BioMed-RoBERTa-RCT had the highest *F*1 score on the validation set and had an *F*1 score of 0.717, precision of 0.689, and recall of 0.748 on test data.

Manual Evaluations

In addition to NLP metrics above, we also manually evaluated results at two points in the project. To assess the viability of application of machine learning to this project and to determine if any improvements could be made to the strategy overall, precision was determined for a 10% random sample of preliminary results mid-way through the project. The second evaluation was performed prior to finalizing the inventory. This evaluation reviewed resources with low probabilities and suspected duplicates. The results of both evaluations are detailed below.

Mid-Project Manual Evaluation

In the mid-project evaluation of a 10% sample, we found that 439/468 (0.938) articles were classified correctly. In Figure 5A, we show an example of a correctly classified article which was relatively straightforward while in Figure 5B, the text describes an entity that appears, on human reading, to be a tool that does not provide access to data [64,65]. Another scenario that proved challenging for the models was when the title-abstract described a research project for which data was deposited into a resource (such as Flybase [66]) but the article did not describe the resource itself. However, overall these errors were limited, and given the challenge of classification even for human curators, we judged that the machine learning based methodology was indeed viable for classification.

A) Correct Positive Classification

PMID 30942868 **Title-Abstract:** Digital expression explorer 2: a repository of uniformly processed RNA sequencing data. Background RNA sequencing (RNA-seq) is an indispensable tool in the study of gene regulation. While the technology has brought with it better transcript coverage and quantification, there remain considerable barriers to entry for the computational biologist to analyse large data sets. There is a real need for a repository of uniformly processed RNA-seq data that is easy to use. Findings To address these obstacles, we developed Digital Expression Explorer 2 (DEE2), a web-based repository of RNA-seq data in the form of gene-level and transcript-level expression counts. DEE2 contains >5.3 trillion assigned reads from 580,000 RNA-seq data sets including species Escherichia coli, yeast, Arabidopsis, worm, fruit fly, zebrafish, rat, mouse, and human. Base-space sequence data downloaded from the National Center for Biotechnology Information Sequence Read Archive underwent quality control prior to transcriptome and genome mapping using open-source tools. Uniform data processing methods ensure consistency across experiments, facilitating fast and reproducible meta-analyses. Conclusions The web interface allows users to quickly identify data sets of interest using accession number and keyword searches. The data can also be accessed programmatically using a specifically designed R package. We demonstrate that DEE2 data are compatible with statistical packages such as edgeR or DESeq. Bulk data are also available for download. DEE2 can be found at http://dee2.io.

B) Incorrect Positive Classification

PMID: 24627222 **Title-Abstract:** miR-Synth: a computational resource for the design of multi-site multi-target synthetic miRNAs. RNAi is a powerful tool for the regulation of gene expression. It is widely and successfully employed in functional studies and is now emerging as a promising therapeutic approach. Several RNAi-based clinical trials suggest encouraging results in the treatment of a variety of diseases, including cancer. Here we present miR-Synth, a computational resource for the design of synthetic microRNAs able to target multiple genes in multiple sites. The proposed strategy constitutes a valid alternative to the use of siRNA, allowing the employment of a fewer number of molecules for the inhibition of multiple targets. This may represent a great advantage in designing therapies for diseases caused by crucial cellular pathways altered by multiple dysregulated genes. The system has been successfully validated on two of the most prominent genes associated to lung cancer, c-MET and Epidermal Growth Factor Receptor (EGFR). (See http://microrna.osumc.edu/mir-synth).

Figure 5. Examples of A) a correctly classified article and B) an incorrectly classified article.

In addition to evaluating the correctness of classifications, in the mid-project evaluation we looked at results of the NER extraction to determine if extracted common and full names were correct, partially correct, or incorrect. A common name was predicted for 457 of the 468 in the sample, and in 425 of these predictions, the common name with the highest probability was, in fact, the correct common name. Notably, even when not correct, only 10 were entirely incorrect (e.g. predicted as “the” instead of “WikiPathways Database”) while 22 were partially correct (e.g. predicted as “Open” instead of “Open TG-GATEs”). In reality, for several of the ones that were entirely incorrect, the name was so poorly articulated in the abstract that it was difficult for a human to determine a valid name. Full names were predicted less frequently on the whole (157/468) and, reflective of greater complexity (see Figure 6 for an example), were less likely to be judged correct with 97 being correct. However, only 7 were entirely incorrect and 53 were partially correct. From these results, it was clear to us that, as was true with predicted classification, the method was viable for predicting resource names, with common names being the most likely identified.

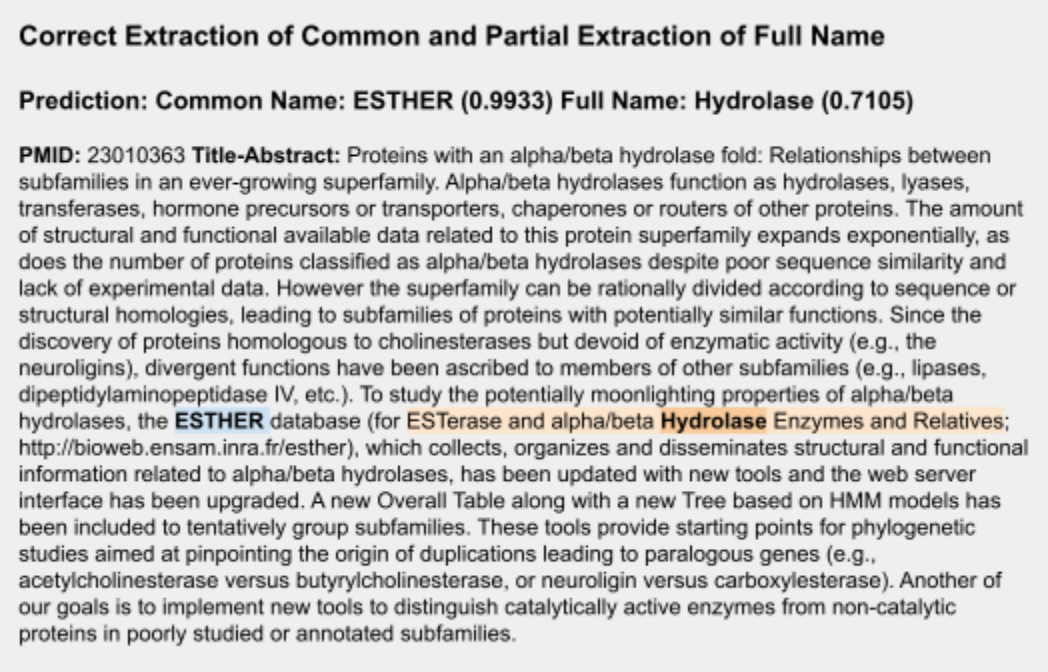

Figure 6. Example of a detailed abstract where the common name of the resource, “ESTER” (bold/blue), was only mentioned once but correctly predicted. The predicted full name, “Hydrolase” (bold/orange), was partially correct while the entire full name (peach) is much longer. The uncertainty is accurately reflected in the scores where the correctly predicted common name is associated with a probability of 0.9933 while the partially correct full name was lower at 0.7105.

While it was clear from the results of the mid-project evaluation that the machine learning methods would be a useful technique to employ, we also recognized that the quality of inventory, as a collection of discrete resources, is important to the stakeholders. We explored putting greater weight on precision than *F*1 for both the classification and NER tasks, but we found that the subsequent hit to recall in the NER task resulted in the loss of too many viable predictions. To achieve higher reliability for the inventory then we choose a selectively mediated approach. Based on the mid-project evaluation results we used the average probability of correctly predicted names, 0.978, to determine a threshold at which resources whose highest scoring predicted name (“best name,” regardless if common or full) had a probability below this threshold would be flagged for review by a curator in our finalized process. In this 10% random sample, records with probabilities ≥ 0.978 (i.e. those that would slip by a human-mediated review) contained 13/468 (2.8%) incorrect classifications and 5/468 (1.0%) incorrect best names. Therefore, we concluded that a selective manual evaluation at this threshold was likely to catch the majority of both classification errors and name errors and improve the quality of the final inventory.

During preparation of this manuscript, we realized that a portion of the ML test sets may have been present in the set of articles that were manually reviewed during the mid-project evaluation. While we did not pass labeled data through the pipeline, we did process all of the articles returned from Europe PMC. The potential impact of such data leakage was assessed. In the mid-project evaluation, 7 articles from the classification test and 5 articles from the NER test set were found in the validation set used for mid-project evaluation. We considered retroactively removing/replacing these 7 articles to

address the potentially deleterious effects of such data leakage but decided against these actions for several reasons. First, these articles made up only a very small portion of the mid-project evaluation set, and we did not look at them specifically, so their effects on decisions regarding ML model training/selection were negligible. Second, the only decision made regarding ML training/selection due to the mid-project evaluation was to use precision rather than *F*1-score for article classification model selection. Due to the very low portion of test set articles in the evaluation (≤1.5%), we are confident that such a decision would have been reached in their absence. Since we believe no portion of the model training or selection process was biased by the presence of these articles, the test sets still serve as representative samples for model evaluation. Third, retroactively removing/replacing these 7 articles would result in convoluting an already complex pipeline and reporting of methods. Ultimately, since our goal was a practical application of ML methods, we opted to simply openly disclose the issue.

Selective Manual Evaluation of Preliminary Inventory

With the results above, the pipeline was implemented and the preliminary inventory up to the point of manual evaluation was created. Of the 3566 records remaining after automated deduplication, 1033 were flagged for review due to a probability score < 0.978, 469 due to potential duplicate names, and 215 due to potential duplicate URLs.

Of the 1033 low probability flags, a curator determined that 805 be retained in the inventory and 228 (6.4% of the total and 22.1% of the flagged) be removed. The majority of those removed, 161/228 (70.6%), were removed because of a partially correct name, while 36/228 (15.8%) were removed for an incorrect name, 27/228 (11.8%) for incorrect classification, and 4/228 (1.8%) for erroneous URLs. As expected, the average probability of those removed (0.749) was much lower than the cut-off that triggered the review.

Of the 469 flagged for duplicate names, 355 were marked for merger, 54 were associated with records that would be removed due to low probability, 50 were not to be merged, and 10 required a partial merger (e.g. sets of > 2 potential duplicates where only some should be merged). Of the 215 flagged for duplicate URLs, 121 were marked for merger, 66 were associated with records that would be removed due to low probability, and 28 were not to be merged.

URL Testing

The Python requests package was used to check the HTTP statuses of the extracted URLs on 11 November 2022. The HTTP status code (e.g., 200 OK) was recorded. In the final inventory 2235/3112 (71.8%) resources had at least one URL which returned HTTP codes in the “2xx successful” or “3xx redirection” series, indicating a live site. However, we note that URLs that resolve with either 2xx or 3xx status codes can be misleading since a URL might resolve successfully but not actually point to the biodata resource. Analysis of 2264 resolving URLs that were manually reviewed in an earlier study of biodata resources revealed that 147/2264 (6.5%) did not direct to the resources itself and instead lead to pages such as a discontinued notice or or generic university page (see Supplemental Methods) [67]. We anticipate a similar false positive rate would apply to this work as well. Additionally, 2451/3112 (78.8%) of resources had at least one URL archived in the Internet Archive’s WayBack Machine.

Assessing the Biodata Resource Inventory

After curation and processing to remove erroneous predictions and deduplicate records that were identified in the selective manual evaluation, the resulting inventory contained 3112 resources from 3705 unique articles. Post processing was completed to add additional metadata and check URLs. We provide the following key descriptive statistics to highlight how the inventory may be used to probe for additional information about the resources.

Countries Represented

Resource locations were assessed by geolocating the URL host IP address and by mining the author affiliations. For 1672 (53.7%) resources at least one IP address could be geolocated, with a total of 1679 IP address locations (Figure 7A). When geolocating IP addresses both coordinates and country name were returned. ISO-3166-1 Alpha-3 codes and country names of all countries were searched against the IP address geolocations and author affiliations [68]. While this yielded a consistent output, it misses certain countries, especially for author affiliations where locations are reported more idiosyncratically. For instance, the ISO-3166-1 name for South Korea is “The Republic of Korea”, and if the name does not appear as such in the affiliations, it is missed. Additionally, false positives may occur, such as an affiliation with New Mexico (a state in the USA) being counted as Mexico (the country). However, as a preliminary assessment, 65 unique countries were found in author affiliations (Figure S4) and 28 unique countries were found in the host IP address locations (Figure S3). Despite the challenges with cleanly identifying countries, over twice as many countries were identified via author affiliations, indicating that global contributions go well beyond the discrete physical location of a resource. This further highlights the distributed nature of the overall infrastructure.

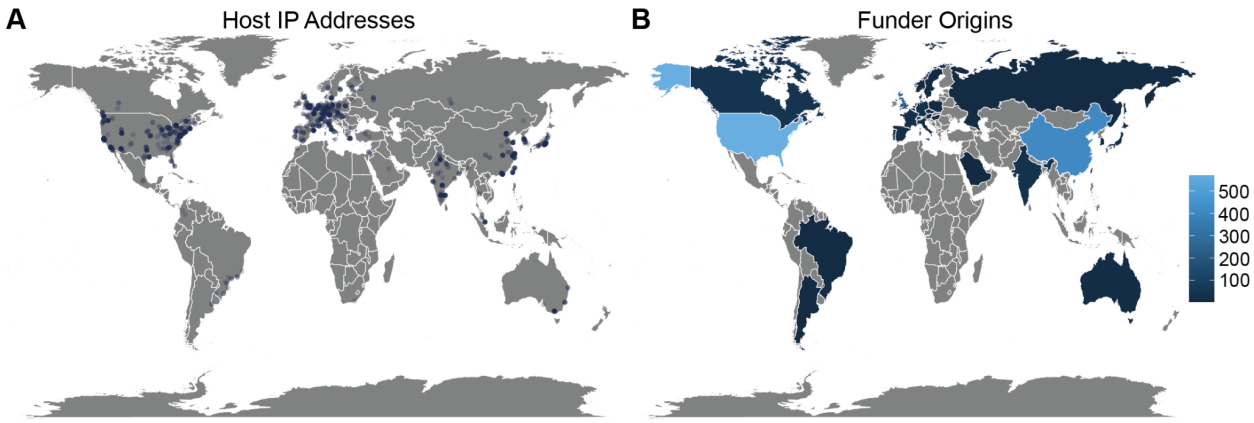

Figure 7. A) URL host IP address coordinates (n=1679). Transparency is constant for all points; darker regions indicate overlapping points. B) Number of biodata resources per country after determining the country of origin for the most frequently identified funding agencies. Funding agencies that could not be mapped to a single country are not shown; specifically, 111 resources were funded by international agencies.

Metadata for Follow-up Analyses of Funder and Text Mining

Funding of biodata resources can be obtained from the granting agency metadata provided for individual articles. Of the 3705 articles, one or more funding agencies were retrieved for 1916 articles

(52.9%), which covers 1714/3112 (55.1%) of the biodata resources identified. Although a total of 1788 “unique” agency names were retrieved, the names reported are variable in that they are often free text values provided by article authors. Both PubMed Central and Europe PMC, which exchange data, make efforts to standardize funder information for their parent funders; however, smaller funders and funders outside of the US and Europe are more likely to be represented irregularly. This challenge is exacerbated by the reporting of increasingly diverse and granular funders as the number of associated biodata resources decreases to only 1 or 2, where reported funder names may be very specific (e.g. an academic department) yet vague (e.g. ambiguous with respect to which university), related to an individual (e.g. scholarships or fellowships of unclear origin), or attributed to a project which, on investigation, is found to be funded itself by multiple funders. Additionally, the funder names retrieved may be of varying organizational levels (e.g. both NIH as a parent and NIH NIGMS as a child), and the same funder can be cited in a single article for distinct grants. In this first instantiation of the inventory itself we left all agency names as is, but future iterations of the inventory may explore use of funder registries, such as the one being developed by CrossRef, to standardize funder names and map identifiers, where possible [69].

Yet preliminary analyses are again illuminating and immediately show the global nature of the infrastructure. Even just a casual scan of funders revealed national organizations from around the world, including but not limited to, the Research Council of Norway, the Spain Ministry of Science and Innovation, the Czech Science Foundation, South Africa’s National Research Foundation, the Israel Science Foundation, the Qatar National Research Fund, India's Department of Health Research, the National Research Foundation of Korea, the Ministry of Science of Technology of Taiwan, the Australian Research Council, Argentina's National Agency for the Promotion of Research, Technological Development and Innovation, the Mexican National Council of Science and Technology, and the Oneida Nation Foundation. To begin assessing the global distribution of the most prevalent funders found here, the funder names associated with 3 or more biodata resources (200/1788, 11.2%) were evaluated by a curator to identify the country or region of the funding organization and verified by a second curator. As noted in prior work, funders themselves are diverse [70], and in addition to government agencies, academic and philanthropic organizations were also identified here. After deduplication of resources, 28 unique counties were identified, with each supporting anywhere from 3 to 570 biodata resources (Figure 7B).

The articles associated with the biodata resources within the inventory also provide opportunities to glean additional information about the resources themselves through text mining. For each article, Europe PMC provides in-house text mining results as annotations that cover database accessions and resource names of over 60 biodata resources as well as gene/protein names, organisms, diseases, chemicals, gene ontology terms and experimental methods [17]. Additional annotations submitted from the community cover specific interactions, targets, pathways, processes, and other terms [71]. Europe PMC offers a dedicated Annotations API for access to all available terms, and new text mining may also be undertaken on full text available via XML. Thus the vast majority of associated articles have text-mined terms already available for further analyses and the majority are available for additional full text analysis (Figure 8).

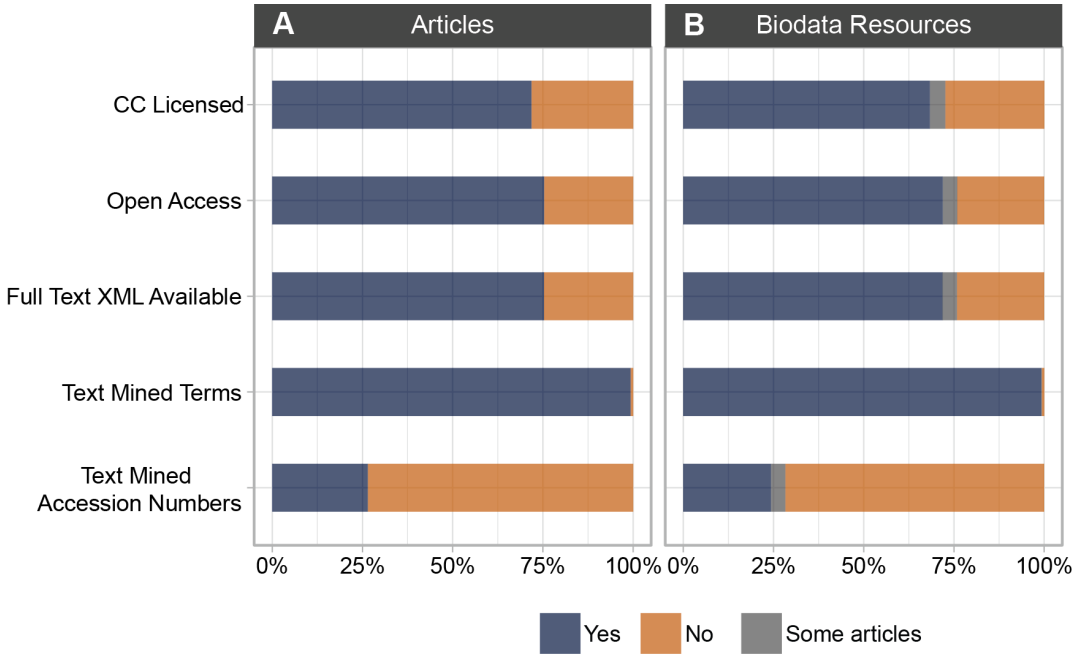

Figure 8. Text mining potential of A) individual articles found to describe biodata resources, and B) the collective articles describing a particular biodata resource.

Overlap with Resources Found in re3data.org and FAIRsharing

The machine learning-enabled approach used here represents an effort to identify resources at scale. Our hope was that this method would be able to augment existing efforts, such as those reviewed in the introduction above. Many of those, including the catalogue maintained by the journal *Nucleic Acids Research* and Database Commons, do not offer bulk download or programmatic access to their data. As such, we focused on two resources, re3data.org and FAIRsharing, which carefully curate their registered resources and provide API access. To investigate the overlap between the inventory we compared the resources identified in the inventory with life science data resources found in both registries. Both resource names (common and, when available, full) and URLs were compared since a resource may have multiple names or multiple URLs as described above. We initially expected greater overlap, but only 536/3112 (17.2%) inventory resources were identified in these two registries; similarly, the majority of life science resources within both re3data (975/1189, 80.5%) and FAIRsharing (1161/1640, 70.8%) are not found in our inventory. It is possible that not all resources in FAIRsharing meet our criteria for inclusion (or vice versa) or that some matches have been missed due to name/URL variations, but at face value these results suggest an even greater complementarity of methods than originally anticipated. Additionally, because this inventory is based on associated articles that describe a given biodata resource, there is an opportunity to reach out to the corresponding authors of the 2609 resources in the inventory without apparent representation in either registry to encourage registration in order to increase the discoverability of their resources. While we designed this project assuming that updating the inventory will be necessary, the ideal future state is that all biodata resources are easily located elsewhere.

Open Science Products

From the onset, our goal was to create a fully open project with products that could be reused by anyone and to a high level of reproducibility—we consider this an explicit result of the project and describe these products here briefly (see Table S7 for listing and locations). The final inventory itself, along with all trained models, code, and data are available in both GitHub (immediately) and in Zenodo (at the conclusion of peer review). Within these deposits is the entire workflow, which is provided both as individual scripts as well as an automated Snakemake pipeline that allows execution of the entire analysis in a single command.

In addition to all of the code, iPython notebooks (ipynb) have also been created to facilitate 1) running the entire pipeline with model training and 2) running the prediction script itself from the highest scoring model to update the inventory. Since running the entire pipeline is computationally intensive and requires access to GPUs, the second notebook may be especially useful for updating the inventory without testing or retraining new models. Detailed READMEs for all products are available within GitHub and Zenodo, and a dedicated protocol has been created to provide step-by-step instructions for use of the iPython notebooks in Google Colab, which allows anyone to execute the code in a browser.

Discussion

Collectively, biodata resources form a life sciences infrastructure that is widely distributed and difficult to describe [72]. While there are a number of highly visible and firmly established biodata resources, others are relatively small and serve specific research communities. In fact, individual resources vary wildly and a persistent, major challenge is simply knowing what biodata resources exist. Worldwide, several national or regional funders, for example in China, Europe, and the US, support institutes that provide major biodata resources, such as those that host gene sequences and protein structures. Such resources are well-known and easy to locate. However, a biodata resource can be established by anyone who wishes to share data and has access to the skills and technical infrastructure needed to create an online resource; that is, the barrier to entry is relatively low and many resources have been created by individual researchers motivated to share data. While this increases the availability of data and aligns with global efforts to increase research data sharing, such resources have proliferated to a point where it is no longer possible to know exactly how many resources exist [73]. This creates challenges for anyone searching for data as well as those who work to develop, maintain, and sustain the resources, including resource providers and research funders.

In an effort to address this issue, we implemented a method to identify thousands of biodata resources via a machine learning enabled pipeline. In this practical application of NLP techniques we found that several state-of-the-art BERT models performed well in both article classification and NER tasks. This paper presents preliminary efforts to provide a comprehensive proof-of-concept, and additional work is already underway to further optimize the ML results. Even as preliminary work, we found these tasks performed well enough that the error rate could have been considered acceptable without remediation. While we ultimately decided to augment with a selective, human-mediated review to further enhance confidence and veracity in the resulting inventory, we conclude that the application of NLP models is a powerful tool that dramatically aids completion of what would otherwise be an

entirely manual, time intensive, and less systematic process. We believe that this work provides a useful approach for addressing a major challenge in sustaining biodata resources—simply being able to monitor an ever-evolving, distributed infrastructure.

Previous and current efforts to create collections of data resources tend to rely on individual effort for discovery and curation. The results of such work are high quality and have succeeded in enabling the discoverability of thousands of distinct resources. This inventory does not seek to replace such highly curated collections but does complement them. The main emphasis of this project was to gather together as many resources as possible using a single, reproducible method. We note, however, that many biodata resources were not identified through our pipeline for several reasons, including 1) biodata resources for which there are no published descriptions, 2) biodata resources described in articles that are not indexed in Europe PMC, and and 3) biodata resources for which descriptive articles have been published but that were missed using our methodology through exclusion, misclassification, or inability to extract a resource name.

While we initially expected to find many of the resources in existing registries, we were surprised by the low overlap which suggests that these methods are even more complementary than we anticipated, with each strategy illuminating largely different sets of resources. More work must be done in order to understand what drives this distinction. For example, it might stem from a difference in the inclusion criteria or the discovery method itself. Regardless, one potential pathway for synergy is using the inventory as a catalyst for outreach to, for example, journals found to frequently publish resource articles or even the authors themselves could be contacted to encourage more submission of biodata resources in relevant registries.

At the onset of this project we identified openness and reproducibility as a key criterion in order both to ensure that the inventory can be updated periodically and so that other researchers are able to reuse and extend the inventory. A preliminary analysis of the inventory indicates that the availability of article metadata and the high percentage of full text articles will indeed enable reuse. In spite of inherent limitations, for example with irregular representation of funding organizations and locations in the associated metadata, the inventory already provides a useful glimpse into this difficult-to-describe distributed infrastructure. Given the global importance of biodata resources, individually and as a distributed infrastructure, we hope that our efforts to make the inventory completely open and to develop the code in ways that make it accessible to the widest possible audience will help catalyze future work in understanding the global infrastructure of biodata resources.

Data Availability Statement

All data and code are openly available, along with documentation, in both GitHub (https://github.com/globalbiodata/inventory_2022) and Zenodo (DOI tbd; archival copy to support article will be deposited post peer review). Please see Supplemental Table S7 for listing of all product types and their locations.

Author Contributions

HJI - Project conceptualization and planning; Manual data curation; Development of preliminary code for data analysis; Project oversight and administration; Data validation; Writing of manuscript
KES - Manual data curation; Implementation of machine learning methodology; Design and implementation of code for processing and augmenting predicted resources, automation of pipelines, unit testing and static code checks, and data analysis; Creation of data visualizations and figures; Data validation; Writing of manuscript
AMI - Design of the machine learning methodology; Implementation of code for training, prediction and evaluation of the NLP models used to classify articles and extract individual resources; Writing of manuscript
CEC - Project conceptualization and planning; Manual data curation; Funding acquisition; Project oversight; Data validation; Writing of manuscript

Acknowledgements

The authors would like to thank colleagues at the Chan Zuckerberg Initiative, in particular Dario Taraborelli, Donghui Li, Gully Burns, and Emanuele Bezzi, for their support and feedback on earlier versions of this study. We also thank Ken Youens-Clark formerly at The University of Arizona, Alise Ponsero at The University of Helsinki, and Bonnie Hurwitz at The University of Arizona for their mentorship of Kenneth Schackart. Additionally, we thank the Europe PMC team, especially Aravind Venkatesan, Mohamed Selim, and Melissa Harrison, for their guidance and expertise. Finally, we would like to acknowledge Jodie Forbes for detailed review of the associated code and documentation.

Funding

This work was supported by the Chan Zuckerberg Initiative, which is a member of the Global Biodata Coalition. This work was also funded by the Global Biodata Coalition (globalbiodata.org), a coalition of research funding organizations working towards sustainability of biodata resources worldwide.

References

1. Drysdale R, Cook CE, Petryszak R, Baillie-Gerritsen V, Barlow M, Gasteiger E, et al. The ELIXIR Core Data Resources: fundamental infrastructure for the life sciences. Bioinformatics. 2020;36: 2636–2642. doi:10.1093/bioinformatics/btz959
2. Gabella C, Durinx C, Appel R. Funding knowledgebases: Towards a sustainable funding model for the UniProt use case. F1000Research; 2018. doi:10.12688/f1000research.12989.2
3. Southan C, Cameron G. D2.1: Database Provider Survey report for ELIXIR Work Package 2. Zenodo. 2017 [cited 2 Jan 2018]. doi:10.5281/zenodo.576013
4. Bourne PE, Lorsch JR, Green ED. Perspective: Sustaining the big-data ecosystem. Nature. 2015;527: S16–S17. doi:10.1038/527S16a
5. Anderson WP. Data management: A global coalition to sustain core data. Nature. 2017;543: 179–179. doi:10.1038/543179a

6. Discala C, Benigni X, Barillot E, Vaysseix G. DBcat: a catalog of 500 biological databases. Nucleic Acids Research. 2000;28: 8–9. doi:10.1093/nar/28.1.8
7. Blair J, Gwiazdowski R, Borrelli A, Hotchkiss M, Park C, Perrett G, et al. Towards a catalogue of biodiversity databases: An ontological case study. Biodiversity Data Journal. 2020;8: e32765. doi:10.3897/BDJ.8.e32765
8. National Institutes of Health. Open Domain-Specific Data Sharing Repositories. [cited 29 Jun 2022]. Available: https://web.archive.org/web/20220629130906/https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html
9. New Mexico State University. Finding Data Repositories. [cited 3 Jan 2023]. Available: https://web.archive.org/web/20230103221358/https://nmsu.libguides.com/c.php?g=400282&p=2901830#Biology
10. PLOS One. Recommended Repositories. [cited 27 Oct 2022]. Available: https://web.archive.org/web/20221027180613/https://journals.plos.org/plosone/s/recommended-repositories
11. Wikipedia. List of biological databases. Available: https://web.archive.org/web/20220901083649/https://en.wikipedia.org/wiki/List_of_biological_databases
12. Rigden DJ, Fernández XM. The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection. Nucleic Acids Research. 2023;51: D1–D8. doi:10.1093/nar/gkac1186
13. Pampel H, Vierkant P, Scholze F, Bertelmann R, Kindling M, Klump J, et al. Making Research Data Repositories Visible: The re3data.org Registry. PLOS ONE. 2013;8: e78080. doi:10.1371/journal.pone.0078080
14. Sansone S-A, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, et al. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol. 2019;37: 358–367. doi:10.1038/s41587-019-0080-8
15. Bandrowski A, Brush M, Grethe JS, Haendel MA, Kennedy DN, Hill S, et al. The Resource Identification Initiative: A cultural shift in publishing. F1000Research; 2015. doi:10.12688/f1000research.6555.2
16. Wren JD, Georgescu C, Giles CB, Hennessey J. Use it or lose it: citations predict the continued online availability of published bioinformatics resources. Nucleic Acids Res. 2017;45: 3627–3633. doi:10.1093/nar/gkx182
17. Ferguson C, Araújo D, Faulk L, Gou Y, Hamelers A, Huang Z, et al. Europe PMC in 2020. Nucleic Acids Research. 2021;49: D1507–D1514. doi:10.1093/nar/gkaa994
18. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28: 2520–2522. doi:10.1093/bioinformatics/bts480
19. Imker HJ, Schackart KE. Open Science Implementation Plan for the Biodata Resource Inventory. 2022 [cited 2 Dec 2022]. doi:10.5281/zenodo.7392518
20. Heil BJ, Hoffman MM, Markowetz F, Lee S-I, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18: 1132–1135. doi:10.1038/s41592-021-01256-7
21. Schackart III KE, Imker HJ, Cook CE. Detailed Implementation of a Reproducible Machine Learning-Enabled Workflow. Zenodo; 2023 Mar. doi:10.5281/zenodo.7767793
22. Chandra RV, Varanasi BS. Python requests essentials. Packt Publishing Ltd; 2015.
23. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs]. 2019 [cited 25 Apr 2022]. Available: http://arxiv.org/abs/1810.04805
24. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language

representation model for biomedical text mining. Bioinformatics. 2019; btz682. doi:10.1093/bioinformatics/btz682
25. Kanakarajan K raj, Kundumani B, Sankarasubbu M. BioELECTRA:Pretrained Biomedical text Encoder using Discriminators. Proceedings of the 20th Workshop on Biomedical Language Processing. Online: Association for Computational Linguistics; 2021. pp. 143–154. doi:10.18653/v1/2021.bionlp-1.16
26. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. arXiv:200410964 [cs]. 2020 [cited 4 May 2022]. Available: http://arxiv.org/abs/2004.10964
27. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv; 2019. Available: http://arxiv.org/abs/1906.05474
28. Miolo G, Mantoan G, Orsenigo C. ELECTRAMed: a new pre-trained language representation model for biomedical NLP. arXiv:210409585 [cs]. 2021 [cited 22 Apr 2022]. Available: http://arxiv.org/abs/2104.09585
29. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthcare. 2021;3: 2:1-2:23. doi:10.1145/3458754
30. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-Alignment Pretraining for Biomedical Entity Representations. arXiv; 2021. doi:10.48550/arXiv.2010.11784
31. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. arXiv; 2019. doi:10.48550/arXiv.1903.10676
32. Ramshaw LA, Marcus MP. Text Chunking using Transformation-Based Learning. arXiv; 1995. doi:10.48550/arXiv.cmp-lg/9505040
33. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics; 2020. pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6
34. Lhoest Q, Villanova del Moral A, Jernite Y, Thakur A, von Platen P, Patil S, et al. Datasets: A Community Library for Natural Language Processing. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. pp. 175–184. Available: https://aclanthology.org/2021.emnlp-demo.21
35. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.; 2009.
36. Harris CR, Millman KJ, Walt SJ van der, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585: 357–362. doi:10.1038/s41586-020-2649-2
37. The pandas development team. Pandas. Zenodo; 2020. doi:10.5281/zenodo.3509134
38. Krekel H. pytest: The pytest framework makes it easy to write small tests, yet scales to support complex functional testing. 2004. Available: https://github.com/pytest-dev/pytest
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12: 2825–2830.
40. PyTorch Team. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019. Available: https://pytorch.org
41. Costa-Luis C da, Larroque SK, Altendorf K, Mary H, Sheridan R, Korobov M, et al. tqdm: A fast, Extensible Progress Bar for Python and CLI. Zenodo; 2022. doi:10.5281/zenodo.7046742
42. Petrov A, Larson SM, Verma P, Garg H. urllib3: Python HTTP library with thread-safe connection pooling, file post support, user friendly, and more. 2008. Available: https://github.com/urllib3/urllib3

43. Niu J. An Overview of Web Archiving. D-Lib Magazine. 2012;18. doi:10.1045/march2012-niu1
44. Davis TL. argparse: Command Line Optional and Positional Argument Parser. 2022. Available: https://CRAN.R-project.org/package=argparse
45. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. 2022. Available: https://CRAN.R-project.org/package=dplyr
46. Jahn N. europepmc: R Interface to the Europe PubMed Central RESTful Web Service. 2021. Available: https://CRAN.R-project.org/package=europepmc
47. Wickham H. forcats: Tools for Working with Categorical Variables (Factors). 2022. Available: https://CRAN.R-project.org/package=forcats
48. Kahle D, Wickham H. ggmap: Spatial Visualization with ggplot2. The R Journal. 2013;5: 144–161.
49. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. Available: https://ggplot2.tidyverse.org
50. Hester J, Bryan J. glue: Interpreted String Literals. 2022. Available: https://CRAN.R-project.org/package=glue
51. Iannone R, Cheng J, Schloerke B, Hughes E, Seo J. gt: Easily Create Presentation-Ready Display Tables. 2022. Available: https://CRAN.R-project.org/package=gt
52. Wickham H. httr: Tools for Working with URLs and HTTP. 2022. Available: https://CRAN.R-project.org/package=httr
53. Ooms J. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:14032805 [statCO]. 2014. Available: https://arxiv.org/abs/1403.2805
54. Bache SM, Wickham H. magrittr: A Forward-Pipe Operator for R. 2022. Available: https://CRAN.R-project.org/package=magrittr
55. Brownrigg R, Minka TP, Deckmyn A, Becker RA, Wilks AR. maps: Draw Geographical Maps. 2021. Available: https://CRAN.R-project.org/package=maps
56. Henry L, Wickham H. purrr: Functional Programming Tools. 2020. Available: https://CRAN.R-project.org/package=purrr
57. Neuwirth E. RColorBrewer: ColorBrewer Palettes. 2022. Available: https://CRAN.R-project.org/package=RColorBrewer
58. Wickham H, Hester J, Bryan J. readr: Read Rectangular Text Data. 2022. Available: https://CRAN.R-project.org/package=readr
59. Wickham H, Seidel D. scales: Scale Functions for Visualization. 2022. Available: https://CRAN.R-project.org/package=scales
60. Wickham H. stringr: Simple, Consistent Wrappers for Common String Operations. 2019. Available: https://CRAN.R-project.org/package=stringr
61. Müller K, Wickham H. tibble: Simple Data Frames. 2022. Available: https://CRAN.R-project.org/package=tibble
62. Wickham H, Girlich M. tidyr: Tidy Messy Data. 2022. Available: https://CRAN.R-project.org/package=tidyr
63. Wickham H, Hester J, Ooms J. xml2: Parse XML. 2021. Available: https://CRAN.R-project.org/package=xml2
64. Laganà A, Acunzo M, Romano G, Pulvirenti A, Veneziano D, Cascione L, et al. miR-Synth: a computational resource for the design of multi-site multi-target synthetic miRNAs. Nucleic Acids Research. 2014;42: 5416–5425. doi:10.1093/nar/gku202
65. Ziemann M, Kaspi A, El-Osta A. Digital expression explorer 2: a repository of uniformly processed RNA sequencing data. GigaScience. 2019;8: giz022. doi:10.1093/gigascience/giz022
66. Drysdale R. FlyBase. In: Dahmann C, editor. Drosophila: Methods and Protocols. Totowa, NJ: Humana Press; 2008. pp. 45–59. doi:10.1007/978-1-59745-583-1_3
67. Imker HJ. 25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and

Maintenance. Frontiers in Research Metrics and Analytics. 2018;3. Available: https://www.frontiersin.org/articles/10.3389/frma.2018.00018
68. Country Codes - ISO 3166. International Organization for Standardization (ISO); Available: https://www.iso.org/iso-3166-country-codes.html
69. CrossRef. Funder Registry. [cited 23 Oct 2022]. Available: https://web.archive.org/web/20221023092819/https://www.crossref.org/documentation/funder-registry/
70. Imker HJ. Who Bears the Burden of Long-Lived Molecular Biology Databases? Data Science Journal. 2020;19: 8. doi:10.5334/dsj-2020-008
71. Europe PMC. Annotations. [cited 30 Dec 2022]. Available: https://web.archive.org/web/20221230133943/https://europepmc.org/Annotations
72. Martin CS, Repo S, Arenas Márquez J, Blomberg N, Lauer KB, Pérez Sitjà X, et al. Demonstrating public value to funders and other stakeholders—the journey of ELIXIR, a virtual and distributed research infrastructure for life science data. Annals of Public and Cooperative Economics. 2021;92: 497–510. doi:10.1111/apce.12328
73. National Academies of Sciences, Engineering, and Medicine. Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs |. Washington, DC: The National Academies Press; 2020. Available: https://doi.org/10.17226/25639