Supplemental Figures

```
(ABSTRACT:(www OR http*) AND ABSTRACT:(data OR resource OR database*))
NOT (TITLE:(retract* OR withdraw* OR erratum)) NOT (ABSTRACT:(retract* OR
withdraw* OR erratum OR github.* OR cran.r OR youtube.com OR bitbucket.org
OR links.lww.com OR osf.io OR bioconductor.org OR annualreviews.org OR
creativecommons.org OR sourceforge.net OR bit.ly OR zenodo OR
onlinelibrary.wiley.com OR proteomecentral.proteomexchange.org/dataset OR
oxfordjournals.org/nar/database OR figshare OR mendeley OR .pdf OR "clinical
trial" OR registration OR "trial registration" OR clinicaltrial OR "registration
number" OR pre-registration OR preregistration)) AND (SRC:(MED OR PMC OR
AGR OR CBA)) AND (FIRST_PDATE:[2011 TO 2021])
```

Figure S1. Europe PMC Query

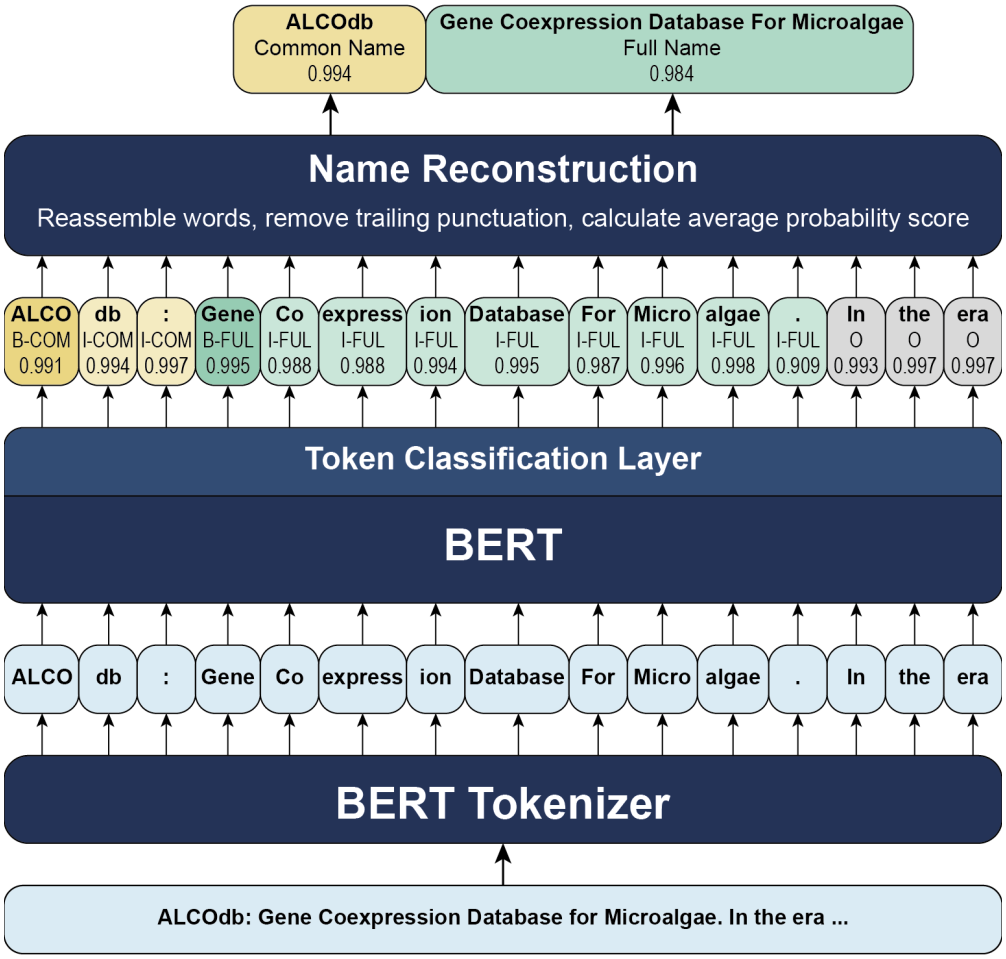

Figure S2. Example of resource name extraction from combined title and abstract. Process shows how the tokens are labeled using the BIO scheme and probability scores are output by the linear token classification layer of the BERT model. Tokens are then reassembled into words using the associated word indices (not shown), and the average probability score of the tokens is calculated. Trailing punctuation is removed from predicted resource names.

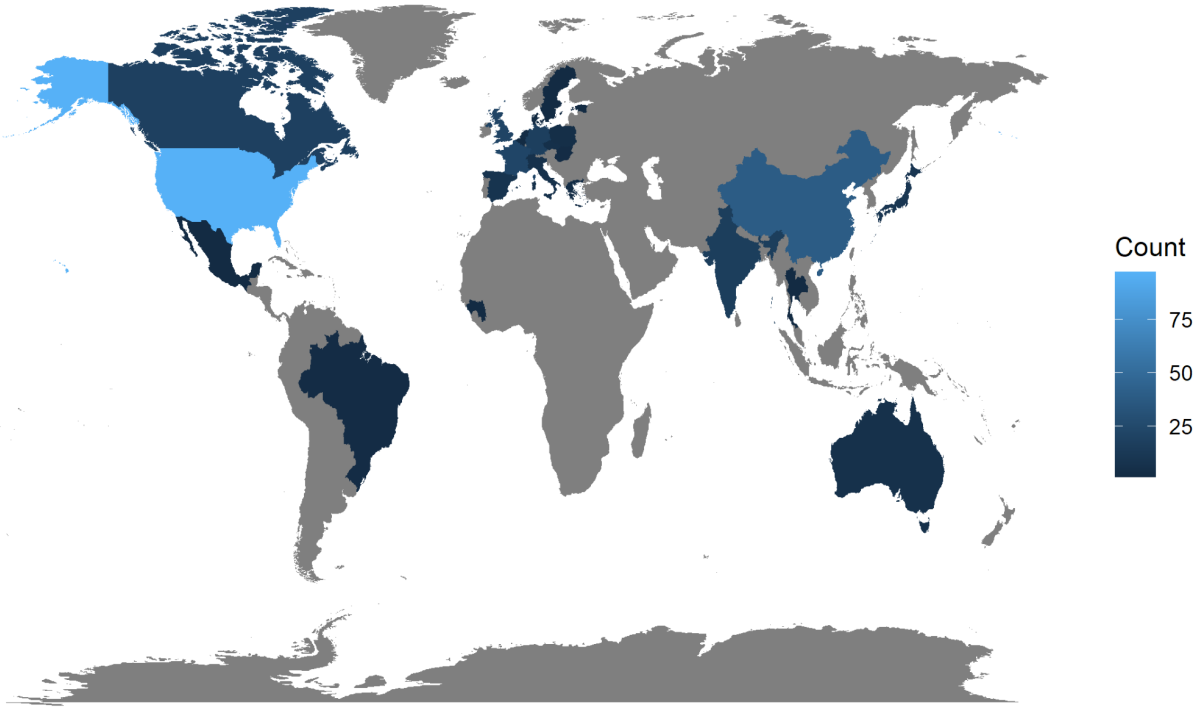

Figure S3. URL host IP address countries based on matches to ISO-3166-1 country names or Alpha-3 codes. Color is scaled to the number of times that country's name appeared as a host IP address location.

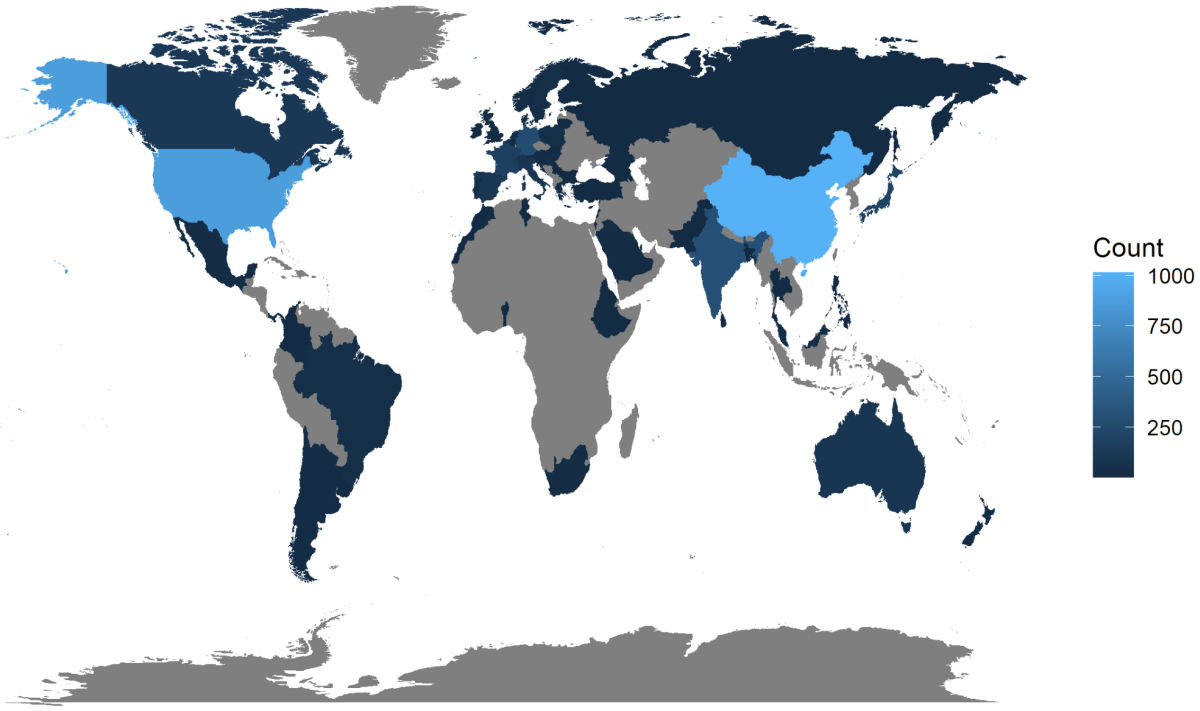

Figure S4. Author affiliation countries based on matches to ISO-3166-1 country names or Alpha-3 codes. Color is scaled to the number of times that country's name appeared in the author affiliations across all articles in the inventory.

Supplemental Tables

Table S1. Definitions Consulted for "Life Sciences Biodata"

Source	**Definition found**	**URL**
Computer Retrieval of Information on Scientific Projects Thesaurus	CRISP:Biology Definition: "science concerned with the phenomena of life and living organisms"	https://web.archive.org/web/20230106212709/https://bioportal.bioontology.org/ontologies/CRISP?p=classes&conceptid=0418-4282
National Cancer Institute Thesaurus	NCIT:Basic Research Definition: "Fundamental research designed to obtain or increase general scientific knowledge."	https://web.archive.org/web/20230106212857/https://bioportal.bioontology.org/ontologies/NCIT?p=classes&conceptid=http%3A%2F%2Fncicb.nci.nih.gov%2Fxml%2Fowl%2FEVS%2FThesaurus.owl%23C15714
Wikipedia	List of Life Sciences	https://web.archive.org/web/

	"branches of science that involve the scientific study of life – such as microorganisms, plants, and animals including human beings"	20211023011543/https://en.wikipedia.org/wiki/List_of_life_sciences
Wikipedia	Basic Research "type of scientific research with the aim of improving scientific theories for better understanding and prediction of natural or other phenomena"	https://web.archive.org/web/20211019010200/https://en.wikipedia.org/wiki/Basic_research

Table S2. Definitions Consulted for "Biodata Resource"

Source	**Definition found**	**URL**
DCAT	dcat:DataService Definition: A collection of operations that provides access to one or more datasets or data processing functions.	https://web.archive.org/web/20221225101619/https://www.w3.org/TR/vocab-dcat-3/#Class:Data_Service
Biomedical Resource Ontology	BRO:Data_Resource Definition: A resource that provides individual facts, statistics or items of information.	https://web.archive.org/web/20230106213856/https://bioportal.bioontology.org/ontologies/BRO/?p=classes&conceptid=http%3A%2F%2Fbioontology.org%2Fontologies%2FBiomedicalResourceOntology.owl%23Data_Resource
DOE	"PuRe Data Resources are data repositories, knowledge bases, and analysis platforms that are sponsored by the Office of Science."	https://web.archive.org/web/20220608144146/https://science.osti.gov/Initiatives/PuRe-Data/Frequently-Asked-Questions
NIH	"... defines data repositories as data resources that store, organize, validate, and make accessible the core data related to a particular system or systems. For example, core data might include genome, transcriptome, and protein sequences for one or more organisms. Knowledgebases are defined as resources that accumulate, organize, and link growing bodies of information related to core datasets. They are resources that may contain, for example, information about gene-expression patterns, splicing variants, localization, and protein-protein interactions and pathway networks related to an organism or set of organisms."	https://web.archive.org/web/20220913053932/https://datascience.nih.gov/sites/default/files/Metrics-Report-2021-Sep15-508.pdf
re3data.org	"A research data repository is a subtype of a sustainable information infrastructure which provides long-term storage and access to research data that is the basis for a	https://web.archive.org/web/20211021235443/https://www.re3data.org/suggest

	scholarly publication. Research data means information objects generated by scholarly projects for example through experiments, measurements, surveys or interviews."	
ELIXIR Core Data Resources	"are of fundamental importance to the wider life-science community and the long-term preservation of biological data. They provide complete collections of generic value to life-science, are considered an authority in their field with respect to one or more characteristics, and show high levels of scientific quality and service"	https://doi.org/10.12688/f1000research.9656.2

Table S3. APIs Used

Source	**Purpose**	**Documentation**
Europe PMC	article corpus	https://web.archive.org/web/20220602202816/https://EuropePMC.org/docs/EBI_Europe_PMC_Web_Service_Reference.pdf
Wayback Machine	URL archive	https://web.archive.org/web/20230106002819/https://archive.org/help/wayback_api.php
re3data.org	resource comparison	https://web.archive.org/web/20230106154142/https://www.re3data.org/api/doc
FAIRsharing	resource comparison	https://fairsharing.org/API_doc
ipinfo	geolocation of IP addresses	https://web.archive.org/web/20230101155943/https://ipinfo.io/developers
ip-api	geolocation of IP addresses	https://web.archive.org/web/20221210042446/https://ip-api.com/docs

Table S4. Hyperparameters used for model training for article classification and NER tasks

Model	**Batch Size**	**Learning Rate**	**Weight Decay**	**Learning Rate Scheduler**
BERT	16	3e-5	0	False
BioBERT	16	3e-5	0	False
BioELECTRA	16	5e-5	0	True

BioELECTRA-PMC	32	5e-5	0	True
BioMed-RoBERTa	16	2e-5	0	False
BioMed-RoBERTa-CP	16	2e-5	0	False
BioMed-RoBERTa-RCT	16	2e-5	0	False
BlueBERT	16	3e-5	0	True
BlueBERT-MIMIC-III	32	3e-5	0	False
ELECTRAMed	16	5e-5	0	True
PubMedBERT	16	3e-5	0	True
PubMedBERT-Full	32	3e-5	0	True
SapBERT	16	2e-5	0.01	False
SapBERT-Mean	32	2e-5	0.01	False
SciBERT	16	3e-5	0	False

Table S5. Article Classification model performance on the validation and test sets, arranged by decreasing precision on the validation set.

	Validation Set			**Test Set**		
Model	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**
BioMed-RoBERTa-RCT[1]	0.849	0.939	0.775	0.821	0.975	0.709
BioMed-RoBERTa-CP	0.800	0.933	0.700	0.791	1.000	0.655
SciBERT	0.800	0.933	0.700	0.791	1.000	0.655
BioBERT	0.783	0.931	0.675	0.821	0.975	0.709
BERT	0.708	0.920	0.575	0.721	1.000	0.564

BioELECTRA-PMC	0.667	0.913	0.525	0.750	1.000	0.600
BioMed-RoBERTa	0.838	0.912	0.775	0.900	1.000	0.818
SapBERT-Mean	0.838	0.912	0.775	0.857	0.977	0.764
BioELECTRA	0.806	0.906	0.725	0.800	0.950	0.691
PubMedBERT	0.806	0.906	0.725	0.854	1.000	0.745
SapBERT[2]	0.886	0.897	0.875	0.874	0.938	0.818
ELECTRAMed	0.846	0.868	0.825	0.871	0.957	0.800
PubMedBERT-Full	0.846	0.868	0.825	0.866	1.000	0.764
BlueBERT	0.773	0.829	0.725	0.860	0.956	0.782
BlueBERT-MIMIC-III	0.773	0.829	0.725	0.832	0.913	0.764

[1] Model with highest precision on validation set that was used to generate final inventory

[2] Model with highest *F*1-score on validation set that was used during mid-project evaluation

Table S6. NER model performance on the validation and test sets, arranged by decreasing *F*1-score on the validation set.

	Validation Set			**Test Set**		
Model	***F*1-score**	**Precision**	**Recall**	***F*1-score**	**Precision**	**Recall**
BioMed-RoBERTa-RCT[1]	0.683	0.674	0.693	0.717	0.689	0.748
BioMed-RoBERTa	0.660	0.685	0.638	0.688	0.681	0.695
SapBERT-Mean	0.651	0.738	0.583	0.651	0.678	0.626
PubMedBERT-Full	0.648	0.717	0.592	0.688	0.736	0.646
SapBERT	0.646	0.733	0.578	0.629	0.674	0.589
SciBERT	0.646	0.680	0.615	0.673	0.656	0.691
BERT	0.644	0.682	0.610	0.703	0.699	0.707

PubMedBERT	0.642	0.620	0.665	0.652	0.638	0.667
BioMed-RoBERTa-CP	0.632	0.646	0.619	0.684	0.686	0.683
BioBERT	0.629	0.644	0.615	0.671	0.693	0.650
BioELECTRA-PMC	0.606	0.649	0.569	0.658	0.675	0.642
BlueBERT	0.606	0.654	0.564	0.643	0.665	0.622
BioELECTRA	0.585	0.598	0.573	0.613	0.605	0.622
BlueBERT-MIMIC-III	0.573	0.608	0.541	0.585	0.567	0.606
ELECTRAMed	0.567	0.647	0.505	0.651	0.670	0.634

[1] Model with highest *F*1-score on validation set that was used during mid-project evaluation and generation of the final inventory

Table S7. Open Science Products

Type	**Name**	**GitHub (living)**	**Zenodo (archive)**	**Hugging Face Hub**	**protocols.io**
data	exact ePMC query output (title-abstract)	https://github.com/globalbiodata/inventory_2022/blob/main/data/epmc_query_results_2022.csv	(to be deposited post peer review of associated article)	n/a	n/a
data	Article classification training data	https://github.com/globalbiodata/inventory_2022/blob/main/data/manual_classifications.csv	(to be deposited post peer review of associated article)	n/a	n/a
data	NER training data	https://github.com/globalbiodata/inventory_2022/blob/main/data/manual_ner_extraction.csv	(to be deposited post peer review of associated article)	n/a	n/a
data	manually reviewed inventory	https://github.com/globalbiodata/inventory_2022/blob/main/data/manually_revie	(to be deposited post peer review of associated article)	n/a	n/a

		wed_inventory.csv			
data	final inventory	https://github.com/globalbiodata/inventory_2022/blob/main/data/final_inventory_2022.csv	(to be deposited post peer review of associated article)	n/a	n/a
code	Python scripts and modules	https://github.com/globalbiodata/inventory_2022/tree/main/src	(to be deposited post peer review of associated article)	n/a	n/a
code	Snakemake workflows	https://github.com/globalbiodata/inventory_2022/tree/main/snakemake	(to be deposited post peer review of associated article)	n/a	n/a
code	ipython notebook for reproducing the original results	https://github.com/globalbiodata/inventory_2022/blob/main/running_pipeline.ipynb	(to be deposited post peer review of associated article)	n/a	n/a
code	ipython notebook for inventory update	https://github.com/globalbiodata/inventory_2022/blob/main/updating_inventory.ipynb	(to be deposited post peer review of associated article)	n/a	n/a
configurations	query	https://github.com/globalbiodata/inventory_2022/blob/main/config/query.txt	(to be deposited post peer review of associated article)	n/a	n/a
configurations	model fine-tuning parameters	https://github.com/globalbiodata/inventory_2022/blob/main/config/models_info.tsv	(to be deposited post peer review of associated article)	n/a	n/a
configurations	Snakemake and directory structure configurations	https://github.com/globalbiodata/inventory_2022/blob/main/config/train_predict.yml	(to be deposited post peer review of associated article)	n/a	n/a
configurations	Python code formatting and	https://github.co	(to be deposited	n/a	n/a

	linting configurations	m/globalbiodata/inventory_2022/blob/main/config/.pylintrc	post peer review of associated article)		
models	fine tuned BERT models	n/a	(to be deposited post peer review of associated article)	https://huggingface.co/globalbiodata/inventory/tree/main	n/a
documentation	README (main)	https://github.com/globalbiodata/inventory_2022/blob/main/README.md	(to be deposited post peer review of associated article)	n/a	n/a
documentation	Google Colab protocol	n/a	n/a	n/a	https://www.protocols.io/view/set-up-biodata-resource-inventory-in-google-colab-5jyl89o36v2w/v1
documentation	open science implementation plan	n/a	https://doi.org/10.5281/zenodo.7392518	n/a	n/a
documentation	curation guide for selective review	n/a	https://doi.org/10.5281/zenodo.7768363	n/a	n/a
documentation	use case article pre-print	n/a	https://doi.org/10.5281/zenodo.7767793	n/a	n/a
documentation	full article pre-print	n/a	https://doi.org/10.5281/zenodo.7768416	n/a	n/a

Supplemental Methods

Analysis of Multi-URL Abstracts

In the mid-project manual evaluation, we found that, although rare, multi-URL abstracts pose a problem. In the 10% sample, 36/468 (7.7%) of abstracts contained 2 URLs and 4/468 (0.85%)contained > 2 URLs with the remaining containing a single URL. In the majority of 2 URL abstracts (26 of 36, ~ 72%), the correct URL was associated with the predicted name and for 6 of the remaining 10, predictions were low enough to trigger manual review, thereby allowing errors to be caught. However, abstracts with > 2 URLs were more often abstracts that covered multiple distinct biodata resources, such as those that describe a collection of resources located at national data centers. As these abstracts were few in number (representing < 1% of the corpus), they were removed instead of undertaking additional training to associate the correctly predicted name with the correct URL.

Analysis of Erroneous Successful URL Status Codes

While creating a census of databases published in *Nucleic Acids Research*, URLs were manually checked and coded when they failed to resolve to the database in question [74]. These data were analyzed to determine how often URLs that appear to resolve successfully do not, in fact, resolve correctly. The variable “unavailable_message” contained two values for failed 2xx codes (“blank page” and “discontinued notice”) and five values for failed 3xx codes ("related generic commercial site redirect", "related generic government site redirect", "related generic publisher site redirect", "related generic research institution site redirect", and "unrelated site redirect"). Of the 2264 URLs that appeared successful, analysis showed that 147 did not resolve properly (6.5%) with 66 resolving to webpages for a related research institution, 37 to an unrelated site, 27 to a discontinued notice, 6 to a related government webpage, 5 to a related publisher webpage, 3 to a related commercial webpage, and 3 to a blank page.

Analysis of Funding Agency Names

To evaluate the funding agency names found in article metadata, the agency names which appeared 3 or more times were extracted and the countries of origin for the funding body were manually identified through Google searches. Recorded country names were verified by a second curator and then standardized to a three letter coding following ISO 3166-1 alpha-3. For funders international in nature, which are particularly prevalent for funding provided through the European Union, a unique 3-letter code of “INT” was used in the absence of an ISO standard.