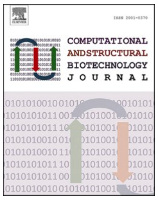
Computational and Structural Biotechnology Journal 21 (2023) 4895–4913

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj



Review article

Metadata integrity in bioinformatics: Bridging the gap between data and knowledge

Aylin Caliskan [a,1], Seema Dangwal [b,1], Thomas Dandekar [a,*,1]

[a] Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany
[b] Stanford Cardiovascular Institute, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305-5101, United States

ARTICLE INFO

Keywords:
Meta-data
Error
Annotation
Error-transfer
Wrong labelling
Patient data
Control group
Tools overview

ABSTRACT

In the fast-evolving landscape of biomedical research, the emergence of big data has presented researchers with extraordinary opportunities to explore biological complexities. In biomedical research, big data imply also a big responsibility. This is not only due to genomics data being sensitive information but also due to genomics data being shared and re-analysed among the scientific community. This saves valuable resources and can even help to find new insights *in silico*. To fully use these opportunities, detailed and correct metadata are imperative. This includes not only the availability of metadata but also their correctness. Metadata integrity serves as a fundamental determinant of research credibility, supporting the reliability and reproducibility of data-driven findings. Ensuring metadata availability, curation, and accuracy are therefore essential for bioinformatic research. Not only must metadata be readily available, but they must also be meticulously curated and ideally error-free. Motivated by an accidental discovery of a critical metadata error in patient data published in two high-impact journals, we aim to raise awareness for the need of correct, complete, and curated metadata. We describe how the metadata error was found, addressed, and present examples for metadata-related challenges in omics research, along with supporting measures, including tools for checking metadata and software to facilitate various steps from data analysis to published research.

1. The effect of reusing data on science

Reusing data can help and accelerate science progress but depends on sound data, accessible raw data and correct metadata. Of course, science does not solely progress due to reusing data, since there are many novel discoveries, findings and insights being made every year, and the technological progress is opening new opportunities. The technological progress and the resulting data, which is rapidly accumulating and getting publicly available for other researchers, can contribute to the progress of science and maybe even accelerate the scientific advancement as researchers have the opportunity to reuse existing data for preliminary analyses, which saves valuable time. The recent COVID-19 pandemic has demonstrated the benefits of next-generation sequencing methods [1], working together [2] and sharing data [3]. That researchers decided to make the initial genome sequence of SARS-CoV-2 accessible for others by uploading it to an open-access database as early as January 2020 was a data-sharing precedent that significantly contributed to subsequent research [3].

However, global pandemics is not the only area of research profiting from publicly available research data. In omics-research, scientists can save time and resources and even cross-check or validate their results by reusing and reanalysing available data. This can speed up research, since it can replace a part of the laboratory research. For instance, analysing existing data can lead to a new research idea, which is already backed by the reanalysis results and might therefore be more likely to prove successful in subsequent laboratory analyses. In 2012, Kodama et al. demonstrated the advantages of this approach by first performing expression-based genome-wide association studies (eGWAS) on already existing microarray data to find a suitable target for treating diabetes: CD44 [4]. In their subsequent laboratory research, Kodama et al. (2012) proved the importance of the immune-cell receptor CD44 in the pathogenesis of diabetes in both mice and humans [4]. This is also indicated by the reuse of publicly available GEO datasets such as the "*Predicting age from the transcriptome of human dermal fibroblasts*" dataset (GSE113957) by Fleischer et al. (2018) [5]. The dataset contains RNA-Sequencing data of human fibroblasts donated by 113 "apparently

[*] Corresponding author.
E-mail address: dandekar@biozentrum.uni-wuerzburg.de (T. Dandekar).
[1] Joined first and senior authors; all authors are equal contributors.

https://doi.org/10.1016/j.csbj.2023.10.006
Received 2 August 2023; Received in revised form 4 October 2023; Accepted 4 October 2023
Available online 5 October 2023

healthy" individuals of all ages between one and 96 years and ten progeria patients. It was published in November of 2018 and reused and cited in 13 PubMed-indexed studies by May 2022 [6]. About a year later (March 2023), the number of studies listed on PubMed that reported the use of the GSE113957 dataset had increased by six. Thus, the dataset has been reused almost 20 times in less than five years, and that is only considering studies listed in PubMed. The number of studies reusing or repurposing the dataset might be significantly larger since it might have been reused in other studies that are either not indexed in PubMed or not published yet.

This demonstrates the growing importance of reliable publicly available datasets in databases such as the Gene Expression Omnibus (GEO) [7,8], which was brought to the forefront of attention a decade ago [9]. Hence, today, in the time of available and re-analysable data, data integrity has become ever more critical, not only for the original researchers' own analyses and results but also for every subsequent analysis that might be performed using the data. An essential aspect of this is the metadata. This additional information can increase the value of the data by adding further details. However, incorrect metadata might have the opposite effect as wrong information in the metadata can lead to inaccurate or even false research results. Therefore, in this review, we aim to raise awareness for the importance of metadata as well as possible metadata errors and their potential consequences and the great responsibility of researchers in ensuring the fidelity of their published data.

2. Metadata and their importance for research

The growing importance of metadata and the need for metadata management in research was already known twenty years ago, as a 2004 review on the evolution, current state and future of metadata by Sen indicates [10]. Additionally, metadata is an integral part of the Semantic Web [11], which was described in a 2001 article in the *Scientific American* by Berners-Lee et al. [11,12] and was envisioned to enhance the World Wide Web by providing machine-understandable information via metadata [13] using the different layers of the semantic web [11,12,14]. Uniform Resource Identifiers (URIs) are metadata and a significant base layer component of the semantic web, which functions similar to international standard book numbers (ISBNs) [11], with Universal Resource Identifiers (URLs) being the most common type of URI [12]. The subsequent layers of the semantic web employ technologies that were already available [12]. The eXtensible Markup Language (XML), which allows adding tags or hidden labels [12], and (in the layer above that layer) Resource Description Framework (RDF), which uses URIs to encode information in triples [12]. These triples are comparable to elementary sentences consisting of a subject, a verb, and an object [12]. This directed, labelled graph data format represents information and metadata. Its specifications define syntax and semantics of the SPARQL Query Language for RDF [15], which was first introduced via a W3C Candidate Recommendation in 2004 and has subsequently been updated several times [16]. The other layers of the Semantic Web also require metadata to enable the agents, which were envisioned to work in a way resembling a personal assistant [12], to function [11]. They include the ontology vocabulary, which has been described in detail by Berners-Lee et al. (2001) [11,12], and allows agents to interpret and use the data. Two decades later, digital assistants, for instance, Alexa and Siri, rely on Semantic Web resources to provide structured content [17]. The core principles of the Semantic Web, such as standardised metadata and ontologies, are also crucial for research. In research data management (RDM), metadata is the foundation for making the data findable, accessible, interoperable and reusable (FAIR) [18]. These criteria, which are also referred to as FAIR Data Principles and FAIRness, have been designed, described and introduced by Wilkinson et al. (2016) as guidelines to enhance the reusability of scholarly data [19] and facilitate sharing, exploring and reusing existing research data [18].

After the original draft, each of the four principles have been refined in the 2016 article introducing the FAIR Guiding Principles by Wilkinson et al. [19], which are also part of the introduction of the FAIR Cookbook by Rocca-Serra et al. (2023) [20], and summarized in Table 1.

The importance of machine-readable (meta)data, which is empathised by the FAIR Guidelines, has also been recognised by other concepts before, for instance by the Semantic Web [21] and is also an aspect of the 5 Star Linked Data Principles [22,23] for Linked (Open) Data, which will be elaborated below.

However, even in 2022, the term "metadata" is not clearly defined, instead a variety of definitions, standards, contexts and formats exists [24]. In fact, according to Furner (2019), there are 96 separate ISO standards and 46 different definitions for the term "metadata" [25]. Additionally, the term is used both for "data about data content", also termed "descriptive metadata", and for "data about data containers", so-called "structural metadata" [25]. Furthermore, it has also been suggested to expand the definition of metadata to the structured and standard part of documentation, and to consider the creation of metadata to the spiral model used in software development and to take the importance of structured and standard documentation during the extended data life cycle into account [26].

A general description could be "data about data" [10,25]. Sen (2004) [10] explains this using the example of measuring the length of a 5 ft stick: in this case, the data is the number 5, while the information on the measurement (what was measured and in which unit?) is regarded as metadata [10]. This perfectly elucidates the importance of metadata, the information "stick" and "5" is of little use without the additional information that the stick's length was measured in ft. The same is true for omics data such as RNA-Sequencing data. Knowing the sample name and the counts of the expressed genes is often not sufficient; most analyses require more information, such as the species or the condition of the sample. Additionally, further information can be of great interest, including age, sex, and – especially for samples derived from human donors – the general health status of the donor and possible comorbidities. Regarding the reuse of omics data, there is a plethora of information that might not be of interest for the researchers who created the dataset, but could be included to facilitate further research.

Table 1
The refined FAIR Guiding Principles, as published by Wilkinson et al. (2016) [19] and in the FAIR Cookbook by Rocca-Serra et al. (2023) [20]. Slightly adapted from Wilkinson et al. (2016) [19].

Findable	Accessible	Interoperable	Reusable
F1. (meta)data are assigned a globally unique and persistent identifier	A1. (meta)data are retrievable by their identifier using a standardised communications protocol	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	R1. (meta)data are richly described with a plurality of accurate and relevant attributes
F2. data are described with rich metadata (defined by R1 below)	A1.1 the protocol is open, free, and universally implementable	I2. (meta)data use vocabularies that follow FAIR principles	R1.1. (meta) data are released with a clear and accessible data usage license
F3. (meta)data clearly and explicitly include the identifier of the data it describes	A1.2 the protocol allows for an authentication and authorization procedure, where necessary	I3. (meta)data include qualified references to other (meta)data	R1.2. (meta) data are associated with detailed provenance
F4. (meta)data are registered or indexed in a searchable resource	A2. metadata are accessible, even when the data are no longer available		R1.3. (meta) data meet domain-relevant community standards

The type of experimental data to be curated also allows ranking the required metadata information on the experiment, allowing dropping of irrelevant information and federating critical information by work flows, masks or data fields.

In order to fully use the potential of publicly available omics data for secondary research, an appropriate annotation is recommended [27]. According to Rajesh et al. (2021), this includes a complete description of the sample type, details on the sample preparation, such as the collection procedure and extraction and assay methods, as well as relevant clinical phenotypes [27]. Additionally, summarized or processed data should be accompanied by metadata containing details about the computational pipeline, for instance the annotation, including which genome build, which gene annotation provenance and which release have been used with which software arguments and versions [27]. The authors also point out that a lack of complete annotations might have a negative impact on follow-up studies intending to reuse the data [27].

Furthermore, improper annotation and incomplete metadata compromise the reproducibility of the original results [27]. Despite the efforts of the biomedical community to share omics data, these efforts are hindered by the lack of consistency among researchers in ensuring the completeness and complete availability of accompanying metadata for raw omics data [27]. Therefore, Rajesh et al. (2021) highlight the need for proper annotation and applying the FAIR principles (Findable, Accessible, Interoperable, Reusable), which have been introduced by Wilkinson et al. in 2016 [19]. They also emphasise the importance of accurate, complete and consistent metadata and a standardised format for both raw data and metadata, which also implies submitting at least a predetermined minimum of clinical phenotypes, including tissue type, age, sex and ancestry [27].

In their assessment of open transcriptomics data across 29 studies, Rajesh et al. (2021) found that on average only 65% of the nine clinical phenotypes they examined were shared publicly, ranging from 83.3% to 38.9% of completeness, with a 35% loss of information between publication and repository accounting for a loss of about 45.7% of the total data between the publication and the corresponding publicly available repository entry [27]. They also stress the importance of rigorous standards for sharing metadata in public repositories to prevent errors caused by the laborious and error-prone approach of scraping metadata from the publication [27]. In summary, metadata needs to be open, complete, freely accessible, standardised and stored in an easy-to-use format to allow other scientists to reproduce the findings of the original publication, enable data reuse, and maximize the utility of the shared data [27].

An important aspect is labelling or sorting the information in the metadata. Ideally, these labels should be standardised to make reusing the data easier. For instance, “tissue type” could be written in several ways, such as “tissue type”, “tissue_type”, and “TissueType” or even “tissue-type” or just “tissue” and these differences might complicate automatic data procession. However, this becomes even more important for terms that are often used interchangeably although they have a difference, such as “race”, “ethnicity”, and “ancestry” [27]. Due to its negative connotation and wrong use it might be best to avoid the term “race”, as Rajesh et al. do in their 2021 publication about omics metadata [27].

This could be implemented by offering predefined metadata categories with defined names and a definition and an explanation regarding the information that is expected in the respective category, for instance like Rajesh et al. (2021) define and explain their use of the term “ancestry”. Additional strategies might be the use of URIs to embed metadata and employing Wordnet synsets containing cognitive synonyms. The Princeton WordNet, which was started in the 1980 s [28], links English words (nouns, verbs, adverbs, and adjectives) to a set of synonyms, which are linked via semantic relations, determining the word definitions [29], and has become a widely used tool in natural language processing (NLP) [28]. Additionally, WordNets have been created for various other languages [28] and Wordnet synsets can even be generated automatically for different languages, including languages with poor resources or endangered languages [30].

Another example of rapidly growing data in need of maintenance and curation are ontologies. By defining concepts, objects and their properties and their relations, ontologies map different types of knowledge and knowledge categories on to the data. Ontologies are used to model scientific fields in order to facilitate computational processing of free text, and to define a vocabulary for standard data formats [31]. As formal representations of ideas, concepts or objects and their relationships, ontologies are often used as controlled vocabularies in requirements engineering during software development [32]. Controlled vocabularies are defined as organized collections of terms with well-known and determined meaning, without duplicates. They facilitate classifying, querying and retrieving data and their usefulness in requirements engineering has been demonstrated [32]. The use of ontologies in the bioinformatics area of proteomics has been described in detail by Mayer et al. (2014), who described the standardised formats and ontologies used in proteomics as well as the ontology formats and appropriate software and the use of controlled vocabularies in the Human Proteome Organisation-Proteomics Standards Initiative (HUPO-PSI) in great detail [31].

However, researchers face challenges when their work requires combining ontologies [33]. This is due to the multitude of different overlapping ontologies, which are used to annotate, organise and analyse data generated by biological experiments and harmonize information of biological knowledge bases but vary in completeness and quality [33]. Hence, the Open Biological and Biomedical Ontologies (OBO) project was founded for organising and guiding the development of ontologies based on shared standards and principles [33]. The Ontology Metadata Vocabulary was created to set metadata standards and to increase the FAIRness of ontology databases, thereby enabling access and reuse of ontologies [33]. All of the ontologies within the OBO Foundry have to fulfil certain requirements and principles, which include shared standards for the interrelation of terms [33]. These principles are stewarded by a team of volunteers that also takes care of various other duties, including metadata curation and maintaining the site [33]. Recently, the OBO Foundry principles were operationalized, and the huge task and the significant community effort involved in re-curating the ontology metadata have been described in detail by Jackson et al. (2021) [33].

Additionally, there are other organizations aiming to provide access to multiple ontologies, such as the National Center for Biomedical Ontology (NCBO), which offers users a uniform mechanism to access a variety of ontologies in different format, including the Open Biological and Biomedical Ontologies (OBO) format and the Web Ontology Language (OWL) format [34].

Other organisations, such as the National Research Data Infrastructure (NFDI, for “*Nationale Forschungsdaten Infrastruktur*”), aim to standardise and harmonise terminologies and identifiers within their infrastructure, for instance by funding initiatives such as the Persistent Identifier Services for the German National Research Data Infrastructure (PID4NFDI) and the Terminology Services 4 NFDI (TS4NFDI) [35]. Additionally, they actively engage with and provide feedback on EU data legislation, such as the EU Data Act, with the objective of refining legal parameters regarding data access, management, and usage to further scientific research and innovation [36]. This aims to increase FAIRness and data accessibility since 80% of industrial data is currently not being used due to various barriers, such as technical, legal and economic barriers [36].

The European life sciences infrastructure ELIXIR even offers a FAIR Cookbook (available at https://faircookbook.elixir-europe.org/content/home.html), an online resource for the Life Sciences offering help and assistance for making and keeping data FAIR [37]. The FAIR Cookbook includes information about the FAIR principles, and various recipes to achieve and optimise Findability, Accessibility, Interoperability, Reusability, Infrastructure, and Assessment [37].

Table 2
Summary of the FAIR Cookbook suggestions for required metadata (modified after chapter "11.5.1 Metadata profile for transcriptomics" of the current FAIR Cookbook (September 2023) [37].

Metadata field	Definition	Comment	Metadata type
unique ID	Identifier for a sample that is at least unique within the project		Common metadata, Assay metadata
sample type	The type of the collected specimen, e.g., tissue biopsy, blood draw or throat swab	ontology field - e.g. OBI or EFO	Common metadata
species	The primary species of the specimen, preferably the taxonomic identifier	This may not be the same as the "host" organism, eg in the case of a PDX tissue sample, the host may be a mouse but the tissue may be human. Ontology field - NCBITaxonomy	Common metadata
tissue/organism part	The tissue from which the sample was taken	ontology field - e.g. Uberon	Common metadata
sex	The biological/genetic sex of the sample	ontology field - e.g. PATO	Common metadata
development stage	The developmental stage of the sample	ontology field - e.g. Uberon or Hsadpdv; species dependent	Common metadata
disease	Any diseases that may affect the sample	This may not necessarily be the same as the host's disease, e.g. healthy brain tissue might be collected from a host with type II diabetes while cirrhotic liver tissue might be collected from an otherwise healthy individual. Ontology field - e.g. MONDO or DO	Common metadata
experiment type	The type of experiment performed, e.g., ATAC-seq or seqFISH	ontology field - e.g. EFO or OBI	Assay metadata
analysis type	The type of analysis performed, e.g., genome assembly or variant calling	ontology field - e.g. EFO, OBI or EDAM	Analysis metadata
platform	The type of instrument used to perform the assay, e.g., Illumina HiSeq 4000 or Fluidigm C1 microfluidics platform	ontology field - e.g. EFO or OBI	Assay metadata
instrument model	The specific instrument on which the assay was performed. Essential for QC purposes.	ontology field - e.g. EFO or OBI	Assay metadata
array or sequencing method	The array or sequencing technology used - may be the same as experiment type or can be a more specific term	ontology field - e.g. EFO or OBI	Assay metadata
extracted nucleic acid/ material type	The type of material that was extracted from the sample, e.g., polyA RNA	ontology field - e.g. ChEBI or EFO	Assay metadata
nucleic acid extraction method	Technique used to extract the nucleic acid from the cell	ontology field - e.g. EFO or OBI	Assay metadata
cDNA library amplication method	Technique used to amplify a cDNA library	ontology field - e.g. EFO or OBI	Assay metadata
end bias	The type of tag or end bias the library has, e.g., 3 prime tag or 5 prime end bias	standardised field or ontology	Assay metadata
biological or technical replicate	Information whether the sample on which the assay was performed was biological or technical replicate.	boolean or CV	Assay metadata
computational method	The specific computational method or algorithm used as part of the analysis	ontology field - e.g. EFO or EDAM	Analysis metadata
normalisation strategy	The approach used to normalise the data	ontology field - e.g. EFO or EDAM	Analysis metadata
file format	The file format in which the analysis is provided	ontology field - e.g. EDAM	Analysis metadata
file storage location	The location in which the data files are stored		Analysis metadata
collection date	The date on which the sample was collected, in a standardised format	Collection date in combination with other fields such as location and disease may be sufficient to de-anonymise a sample	Common metadata

That the additional information is attached to the correct sample is not only important for the original research but also for future research, both for studies reusing the data and for studies citing the results obtained with the data. The possible impact of wrong conclusions and the subsequent multiplication of error as well as the reluctance of publishing results seen as "negative results", or of results that might challenge established practices have both already been eloquently described by Ioannidis (2010) [38].

3. Different types of metadata

There are not only numerous definitions of metadata but also various types of metadata in themselves. The different types of metadata have recently been described by Ulrich et al. (2022), who found 23,233 records for the keyword "metadata" and selected 551 of these records by using suitable keywords and removing duplicates, which were subsequently screened [24]. This resulted in a total 81 records that were subsequently analysed by the researchers [24]. Taking their possible biased selection that resulted in the majority of the papers being from the field of bioinformatics [24] into account, this indicates that defining the term "metadata" is probably even more complicated.

To help researchers decide which information they should or could add as metadata to their experimental data, the FAIR Cookbook contains a recipe for a metadata profile for different types of research data (chapter 11.5.1 of the current FAIR Cookbook (September 2023)) [37]. The FAIR Cookbook contains several extensive but non-exhaustive lists of metadata suggestions for various analyses and differentiates between required and recommended metadata [37]. Table 2 and Table 3 summarise the suggestions for required and recommended metadata, respectively [37], the complete recipe is available at https://faircookbook.elixir-europe.org/content/recipes/interoperability/transcriptomics-metadata.html#assay-metadata.

Among the required metadata are unique identifiers or short URIs (Uniform Resource Identifiers) [37], which are also part of the concept of the Semantic Web. Other required metadata fields include not only the more immediate considerations such as sample type, species and disease but also less intuitive parameters. These can include information whether the sample was a biological or technical replicate for assay metadata, or which computational method or algorithm was employed in the analysis [37].

Although the information of the recommended metadata fields is not strictly necessary for a re-analysis of the data, including as much of the recommended metadata fields as possible can facilitate the re-use of a

Table 3
Summary of the FAIR Cookbook suggestions for recommended metadata (modified after chapter “11.5.1 Metadata profile for transcriptomics” of the current FAIR Cookbook (September 2023) [37].

Metadata field	Definition	Comment	Metadata type
sample collection technique	The technique used to collect the specimen, e.g., blood draw or surgical resection	ontology field - e.g. EFO or OBI	Common metadata
age	Age of the organism from which the sample was collected		Common metadata
age unit	Unit of the value of the age field	ontology field - e.g. UO	Common metadata
ancestry/ethnicity	Ancestry or ethnic group of the individual from which the sample was collected	ontology field - e.g. HANCESTRO	Common metadata
BMI	Body mass index of the individual from which the sample was collected	Only applies to human samples	Common metadata
strain	Strain of the species from which the sample was collected, if applicable	ontology field - e.g. NCBITaxonomy	Common metadata
cell type	The cell type(s) known or selected to be present in the sample	ontology field - e.g. CL	Common metadata
cell location	The cell location from which genetic material was collected (usually either nucleus or mitochondria)	ontology field - e.g. GO	Common metadata
treatment category	Treatments that the sample might have undergone after collection	ontology field - e.g. OBI, NCIt or OGMS	Common metadata
growth conditions	Features relating to the growth and/or maintenance of the sample		Common metadata
genetic variation	Any relevant genetic differences from the specimen or sample to the expected genomic information for this species, e.g., abnormal chromosome counts, major translocations or indels		Common metadata
phenotype	Any relevant (usually abnormal) phenotypes of the specimen or sample	ontology field - e.g. HP or MP; species dependent	
cell cycle	The cell cycle phase of the sample (for synchronized growing cells or a single-cell sample), if known	ontology field - e.g. GO	Common metadata
cell quality	Information about the quality of a single cell such as morphology or percent viability	standardised field or ontology	Assay metadata
cell barcode	Information about the cell identifier barcode used to tag individual cells in single cell sequencing		Assay metadata
UMI barcode	Information about the Unique Molecular Identifier barcodes used to tag DNA fragments		Assay metadata
assay start time	The exact time at which the assay was started		Assay metadata
assay end time	The exact time at which the assay was completed		Assay metadata
assay duration	The duration, in a relevant time unit (e.g., minutes or hours), of the assay from start to finish		Assay metadata
array quality	The overall quality of the array		Assay metadata
chemical compound	Any relevant chemical compounds used in the assay	ontology field - e.g. ChEBI	Assay metadata
labeling molecule used	The type of labeling molecule used in an array-based experiment	ontology field - e.g. ChEBI	Assay metadata
spike-in kit used	Information about the spike-in kit used during sequencing library preparation		Assay metadata
cDNA primer	Type of primer used for cDNA synthesis from RNA, e.g., polyA or random	standardised field or ontology	Assay metadata
library strandedness	The strandedness of the cDNA library	standardised field or ontology	Assay metadata
analysis date	The date on which the analysis was performed		Analysis metadata
read index	The sequencing read a specific file represents, e.g., read1 or index1		Analysis metadata
read length	The length of a sequenced read in this file, in nucleotides.		Analysis metadata
assembly type	The assembly type of the genome reference file, e.g., primary, complete or patch assembly.	standardised field or ontology	Analysis metadata
reference genome version	The genome version of the reference file.		Analysis metadata
software package	The software package used for data analysis		Analysis metadata
software version	The exact version number of the software package		Analysis metadata
external accessions	Accession numbers from any external resources to which the sample was submitted or to which assay or protocol information was submitted	e.g. Biosamples, Biostudies e.g. protocols.io, AE	Common metadata, Assay metadata

dataset and help other researchers to gain more insights when reanalysing the provided data.

4. Data governance

Another important aspect in handling research data and metadata stewardship is data governance. In their 2019 review on data governance Abraham et al. bring to attention that the amount of data that is created is rapidly increasing [39]. The amount of created data was said to increase from 4.4 zettabytes in 2013–44 zettabytes in 2020 [39], which equals 44 trillion (10ˆ12) gigabytes (GB). This is enough storage space for about 6.3 trillion high definition movies of about 7 GB in size or 11 trillion DVDs (4.7 GB). Assuming each song needs about 5 megabytes storage, this would be enough for 8.8 quadrillion hours of music (which is more than one trillion times the current age of the universe, which is estimated to be around 13.8 billion years [40]).

Abraham et al. (2019) also provide a working definition of the term “data governance” as a “*cross-functional framework for managing data as a strategic enterprise asset*”, which also specifies “*decision rights and accountabilities for an organization’s decision making about its data*”, and “*formalizes data policies, standards, and procedures and monitors compliance*” [39]. Thus, the data needs to be managed in a way that maximises its value and manages data-related risks and challenges such as inaccurate and incomplete data and compliance issues need to be overcome, and the conceptual framework has been described in detail by Abraham et al. (2019) [39]. The consequences or outcomes of data governance include a positive effect on data utilisation, increased data quality, and the management of data-related risks due to a better oversight regarding the data quality and risk-mitigating policies to reduce risks concerning privacy or security breaches [39]. Additionally, it has been shown that organizations that are able to use their data effectively, for instance by tagging the data with metadata, have advantages over their competitors [41], which demonstrates the importance of having and handling metadata correctly.

In the healthcare sector, the most important challenges are reliability and integrity, as they are related to life and death [42]. Especially sensitive data such as patient data has to be kept in secure places and should only be accessible to authorized parties, and criminal acts, such as the theft of personal medical history, have to be prevented [42]. Thus, a conflict of interest between collecting and using the data and legitimate concerns arises. In healthcare, data collection, sharing and collaboration face challenges when patient consent is necessary [42]. Therefore, data governance policies need to address privacy, security, and accuracy as well as storage, usage and preservation inside the organization, and data access and lifecycle [42]. Additionally, it is crucial to consider data standards and automation strategies in order to effectively manage data [42]. Legislation of data governance, such as the new EU Data Governance Act, address such topics, including the creation and regulation of so called "secure spaces" for sharing and reusing sensitive data such as health data for commercial and altruistic purposes, which also includes scientific research [43]. An important aspect in sharing biomedical data are access barriers, for instance, the data protection principle of purpose limitation, which states that data can only be used for specific purposes [43]. This hinders the use of the data for multiple research purposes as explicit consent for each downstream use is required [43]. Thus, the data-sharing infrastructure, secure data-sharing platforms and data governance need to be adapted to allow "further processing" and reuse of the data by other scientists [43]. At the same time they need to ensure data protection and privacy, which can be achieved by ensuring that the data is only accessible to authorised users for authorised purposes [43]. A regulatory data governance framework for data-sharing infrastructure can facilitate the sharing of data and thus research [43], and the recent COVID-19 pandemic demonstrated the need for a robust data governance framework [43].

A possible approach to handle big data while its being generated is described by Zimmerman et al. (2014) [44]. They describe how structural genomics centres use mechanisms to connect results into a unified system by employing laboratory information management system (LIMS) tools and central databases, for instance UniTrack, which unifies and curates data obtained by different laboratories [44]. Other tools, such as LabDB, can automatically or semi-automatically harvest data from laboratory equipment [44]. The reagent tracking module of LabDB can track the use of reagents via unique barcodes [44]. When the barcodes are scanned during the preparation of a stock solution, a new unique barcode for the stock solution is created [44]. This barcode allows tracking the origins of the chemicals and carrying detailed information along the pipeline, which provides much more detailed information about the contents of the stock solution than a hand-written label would [44]. Additionally, the data is linked to later steps, which allows determining whether unsuccessful experiments can be traced back to a certain reagent [44]. Systems such as these could also be used to connect metadata about the origin of a sample, e.g., detailed information about the donor, such as their health condition and possible comorbidities, their age, and other information that might be relevant for research. This would allow tracing a samples history back to its origins and connecting this information to further data such as sequencing results. Therefore, automatically uploading sequencing results and their metadata to a database would become much easier, which might help keeping the metadata correct and complete without adding more effort for the researcher.

5. Good data governance

Another example for data quality criteria are the AHIMA characteristics of data quality by the **A**merican **H**ealth **I**nformation **M**anagement **A**ssociation [45], which coincide with the FAIR principles and show how data handling principles such as the FAIR principles might be implemented in the clinic (Fig. 1). The convergence between the FAIR principles and the AHIMA guidelines underscores the widespread recognition of data quality challenges in the field. The AHIMA guidelines can be seen as an essential checklist for practitioners, highlighting the specific qualities imperative for achieving optimal data quality. According to the AHIMA criteria, data needs to be (1) accurate, therefore correct and free of errors, (2) accessible, so that the data is available when required, (3) comprehensive and contain all required elements, (4) consistent, meaning that the data is "reliable and the same across the patient encounter", in terms of sequencing data this term could also be used to describe that every sample of a dataset was prepared according to the same protocol or advise a consistent use of categories for the accompanying metadata, (5) current, which in a clinical setting describes that every information is up to date, could be adapted to emphasize that every step and every additional information should be documented, (6) clearly defined, (7) granular, meaning containing the appropriate level of detail, (8) precise, (9) relevant, which is defined as relevant to the purpose it was collected for, although additional, seemingly not relevant information might be useful for other researchers also using the data, and lastly (10) timely, which the AHIMA defines as entered promptly as well as "up-to-date and available within specified and required time frames" [45], which is a good laboratory practice while generating data and might prevent confusion or even labelling errors.

While unique IDs / URIs are among the required metadata fields and crucial for enhancing the findability of data, it is also important to consider other metadata fields that improve the specific needs of researchers trying to find datasets for specific analyses. During preliminary analyses, researchers might be interested in already available datasets regarding a certain tissue type, a specific disease or a defined age group. Since these search criteria are among the required or recommended metadata fields, an effective search should to include these metadata fields to find suitable datasets and the datasets containing the respective information might easier to find. As rich metadata can enhance the findability of the data, the metadata is an important aspect of data sharing. Additionally, researchers can help other researchers to find their data by adding suitable metadata. Thus, the metadata can affect several aspects of research: In the original research, correct metadata guarantees valid results. In subsequent research, correct (and ideally rich) metadata can affect (1) the findability of the data and the results of database searches, and thus the reuse of the data and (2) the research of others reanalysing the data.

During our data retrieval research for a bioinformatics analysis project, we searched for human lung samples of individuals who were either healthy or infected with SARS-CoV-2. Therefore, for this example, keywords such as "human lung", "lung tissue", "human", "healthy", "infected with SARS-CoV-2″, "SARS-CoV-2″, would have made the data findable. However, samples only tagged with "COVID-19″ or "Corona" would most likely not turn up among the search results, when using only the before mentioned keywords. Keeping this in mind, researchers sharing their data should include as many suitable keywords as possible (e.g., "SARS-CoV-2″, "COVID-19″, "Corona", "novel Corona Virus", …). Additionally, in an ideal world, the search algorithms should be able to find data related these keywords correctly, ideally even if the keywords in the metadata are written slightly differently, contain a typo or are not the actual keyword but a synonym.

This example highlights an important challenge in generating FAIR data: the difficulty of aligning the FAIR principles with the human-centric aspects of data discovery. Musen et al. (2022) underscored that an important aspect of FAIR data is to ensure that metadata encompasses adequate descriptors enabling researchers to find datasets with satisfactory recall and precision [48]. They emphasised the need for machine-processable metadata templates guiding both researchers and data stewards in how community-specific metadata standards should be applied [48]. These templates should contain all necessary community-based details and standards that are required for consistent research metadata. By using such templates, researchers could streamline the process of adding metadata to the respective research data and, at the same time, ensure that the data remains compliant with the FAIR principles [48]. This also highlights the importance of data FAIRification and the effect of metadata on findability.

6. The open data guidelines

The importance of data availability is not only highlighted by the

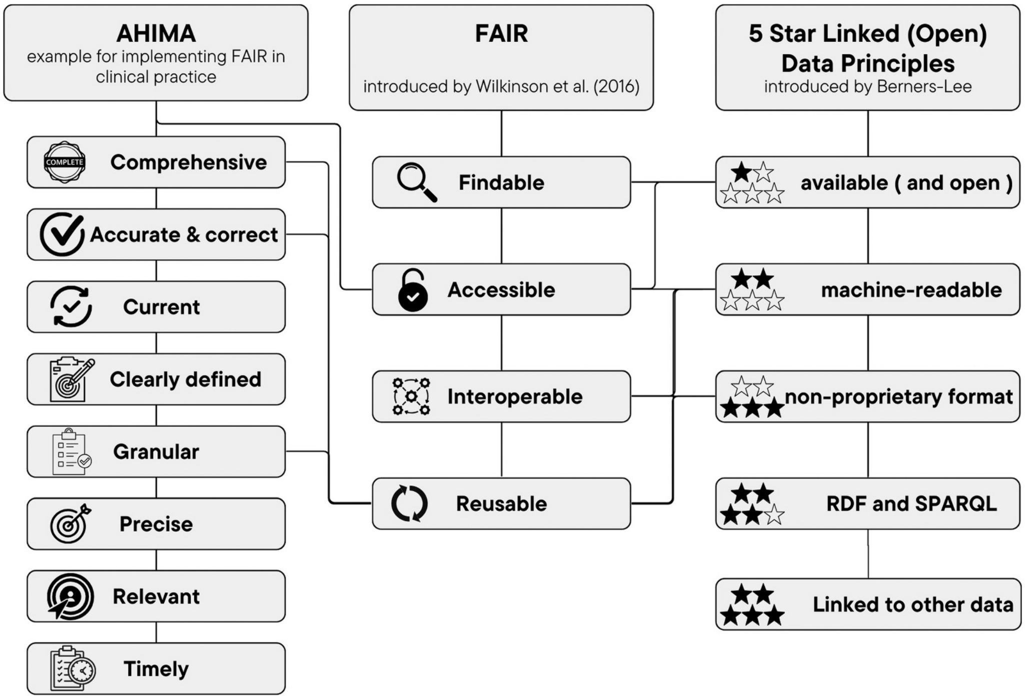


Fig. 1. Schematic representation of the FAIR principles juxtaposed with comparable guidelines such as AHIMA and the 5 Star Linked (Open) Data Principles. FAIR principles = Findability, Accessibility, Interoperability and Reuse of data; AHIMA (American Health Information Management Association) guidelines for optimal data provision in the clinic. 5 Star Linked (Open) Data Principles for step-wise deployment of open data were suggested by internet constructor Tim Berners-Lee. Own figure.

FAIR Principles but also by other concepts, for instance the 'Open Data' principle, which refers to non-confidential and non-private data being made available via public means. Although the FAIR principles and open data share similarities, the concepts are not identical, as Jati et al. (2022) elaborated using Kingdon's multiple streams model [21].

According to the definition by Geiger and von Lucke, Open Data is defined as making "all stored data of the public sector which could be made accessible by the government in the public interest without any restrictions on usage and distribution" accessible [21,49]. The main goals of Open Data is that anyone can freely use, reuse an redistribute the data and the maximization of interoperability [21]. While Open Data aims at providing the public with access to data considered to be in the public interest and excludes confidential, private and classified data, the FAIR Principles were developed in the research environment and focus on challenges in data collection for research [21].

The FAIR aspects Findability and Accessibility are comparable to the aspect of availability in Open Data [21]. Findable data is defined as data that can easily be found by humans and machines, and these data should also be accessible to users, although FAIR does not specify the type of user accessing the data or the type of data being accessed [21]. In Open Data, the data being accessed is required to be non-confidential, as the focus is on making data available to the public. The FAIR Principles, on the other hand, also consider the need for data protection and access requirements. Therefore, FAIR data can be either public or confidential, while Open Data is required to be free from usage restrictions, non-private and non-confidential [21].

Additionally, the FAIR Principles promote interoperability, for instance via machine-readable ontologies and metadata, which can be stored in formats also used in the Semantic Web (e.g., RDF, the Research Description Framework to represent interconnected data on the web) [21]. Open Data does not specifically focus on interoperability, although the concept also emphasizes that "anyone should be able to use, reuse and redistribute the data" [21]. Additionally, the 5 Star Linked (Open) Data Principles emphasise enabling other users, both humans and machines, to utilise the data by advocating for the use of machine-readable, non-proprietary data formats [22,23]. The focus is on the data being available for everyone and the data being reusable for any purpose [21], which can be summarized as redistribution neutrality. However, the structure or format of the data is not explicitly defined in Open Data [21]. This is comparable to the FAIR Principles aspect of Reusability, which additionally encompasses data and metadata being defined for reuse and being able to be replicated or used in different environments [21]. Note that findability of data via metadata is a basis for data retrieval, but in practice different from precision and recall, as changing the metadata (adding more information) will improve the discoverability of the respective data in a database but might not necessarily affect precision and recall of the database search itself. While the results of a database search are affected by precision and recall and the available information (e.g., in the metadata). One has to focus both on the aspect of improving the metadata as well as fast recall and high precision (depending on curation and structure of your database).

While it is possible to achieve the goal of Open Data by applying the first three of the FAIR Principles, FAIR data is not necessarily "open" [21]. FAIR does not aim to make data accessible to the general public but invites data ownership by also considering possible access restrictions due to the nature of the data (e.g., sensitive data) and therefore includes that data users might need to be authorised and verified before accessing the data [21].

6.1. Metadata errors are often only accidentally spotted and difficult to correct

Using two highly cited studies as an example, we intend to raise awareness of the importance of correct metadata and illustrate to the reader why measures to improve software metadata and control the correctness of the data in a timely manner are imperative for good scientific data.

6.2. Finding mistakes in metadata preservation by accident

Mistakes in metadata preservation happen, and their detection is often only possible by accident if datasets are systematically compared. Thus, in our first example on incorrect metadata, we give and review here, that we originally intended to study a scientific question and not the metadata: To find out whether or not alveolar epithelial cells react to SARS-CoV-2, we reanalysed suitable publicly available datasets, including the GEO datasets GSE147507 [50] and GSE155241 [51]. These datasets appeared in flagship publications:

The GSE147507 dataset was generated during the research for the *Cell* publication "*Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19*" by Blanco-Melo et al. (2020) [50] and has been publicly available in GEO [7,8] since March 2020. The authors revealed a unique and inappropriate inflammatory response, defined by low levels of type I and III interferons juxtaposed to elevated chemokines and high expression of IL-6 compared to controls [50]. The other dataset, GSE155241, was generated during the research for the *Nature* publication "*Identification of SARS-CoV-2 inhibitors using lung and colonic organoids*" by Han et al. (2021) [51] and has been publicly available in GEO [7,8] since August 2020. Using alveolar type-II-like cells permissive to SARS-CoV-2 infection, the research group performed a high-throughput screen of approved drugs to identify entry inhibitors of SARS-CoV-2 compared to uninfected controls, such as imatinib, mycophenolic acid, and quinacrine dihydrochloride [51].

6.3. Consequences of metadata errors in high-impact publications

If there is a metadata error, it spreads particularly rapidly if it occurs in high-impact publications: In our example, both high-impact publications have been widely cited: The *Cell* article by Blanco-Melo et al. [50], which was published online on May 15^{th,} 2020, has been cited 2435 times (2425 times in CrossRef, 2417 times in Scopus, and 53 times in PubMed Central (as of April 2023)) according to *Cell*'s PlumX Metrics, and the *Nature* article by Han et al. [51], which was published online on October 28^{th,} 2020 and has been accessed 57k times according to *Nature*'s own metrics, has been cited 246 times in Web of Science and 267 times in CrossRef (as of April 2023).

6.4. Metadata errors can be identified by systematic comparison

For instance, during our analysis of the data from our examples, we found irregularities in the RNA-Sequencing data of both publications (data derived from Blanco-Melo et al. [50] (2020) and Han et al. [51] (2021)): Two of Han et al.'s human lung tissue samples (GSM4697983 and GSM4697984 from GSE155241) [51] appear to be precisely the same as two other unrelated samples of human lung tissue (GSM4462413 and GSM4462414 from GSE147507) that were generated by Blanco-Melo et al. (2020) during research for their *Cell* publication [50]. This became obvious during the first steps of the analysis workflow, which was originally performed in 2020 and has been repeated with newer software and GENCODE versions, still yielding the same results. After preparing the human lung tissue RNA-sequencing data of both publications (GSE147507 [50] and GSE155241 [51]) with STAR alignment (version 2.7.10a) [52], using the comprehensive gene annotation PRI and the genome sequence, primary assembly (GRCh38) PRI of GENCODE version 39 [53,54] and transcript quantification with RSEM (version 1.3.1) [55], the resulting files were analysed in RStudio. After importing the data via tximport (version 1.24.0) [56], we performed a DESeq2 analysis (version 1.36.0 [57], with apeglm version 1.18.0 [58]). In the resulting heatmap (Fig. 2A, generated using the R-package pheatmap, version 1.0.12 [59]) and the resulting principal component analysis (Fig. 2B, using DESeq2 [57])the samples "control_1″ and "control_3″ as well as "control_2″ and "control_4″ appear to express precisely the same genes.

The principal component analysis (PCA) in Fig. 2B confirms and visualizes the similarity between the samples of "Publication_1″ (the samples by Blanco-Melo et al. (2020) [50]) and "Publication _2″ (the samples published by Han et al. (2021) [51]). GSM4462413 and GSM4697983, as well as GSM4462414 and GSM4697984, are superimposed. This is because the respective samples have the same principal component values (see Supplementary Table 1). GSM4462413 and GSM4697983 cluster together and are recognised as one branch of the dendrogram, and GSM4462414 and GSM4697984 cluster together as a branch of the dendrogram, resulting in two dendrogram branches for four samples. The similarity of the results was especially confusing as the respective metadata indicated that sample GSM4462413 (by Blanco-Melo et al. (2020) [50]) was obtained from a male donor while the remarkably similar sample GSM4697983 (by Han et al. (2021) [51]) was obtained from a female donor (according to the metadata and according to personal communication). Additionally, both research groups obtained their samples from different institutions. According to their publication, the Blanco-Melo group obtained their healthy lung tissue samples at Mount Sinai and their SARS-CoV-2 infected lung tissue samples as fixed samples from Weill Cornell Medicine [50]. Han and colleagues obtained all of their tissue samples (control and COVID-19 samples) from the Weill Cornell Medicine Department of Pathology [51]. Analysing the samples revealed that GSM4462413 and GSM4697983 as well as GSM4462414 and GSM4697984 show an identical count and sequencing read distribution, which can also be seen both in the heatmap and the PCA (Fig. 2) and in the visualisation of the sex-specific gene expression (Fig. 3). Additionally, we checked each analysis step as well as the complete pipeline internally (two independent people from our US/German team) to verify that the results were precise and reproducible. Our results indicated that the officially different samples were identical, which has been confirmed and corrected by the respective authors, who updated the GEO-entries: GSM4697983 is now (since January 5th, 2022) labelled as reanalysis of GSM4462413 and GSM4697984 as reanalysis of GSM4462414, respectively. Subsequently, we directly compared the gene expression of the samples in question and contacted the respective authors and journals regarding the striking similarity between the individual samples.

6.5. Objectively identifying metadata on sex

Metadata on sex can be objectively identified in raw data looking at XIST and Y-chromosome specific genes: To demonstrate this again with our example (works, however, for all gene expression data you are interested in to verify correct sex annotated), we compare the sex-specific gene expression of the samples (Fig. 3). Since, according to the metadata, the samples in question were derived from three men and one woman, we compared the expression of *X-inactive specific transcript* (*XIST*), which is responsible for the dosage equivalence of X-linked genes in both sexes and the inactivation of the second X chromosome in females, and thus typically expressed in females [60].

Due to the striking similarity in XIST expression between the two officially male samples, GSM4462414 and GSM4697984, and the complete absence of XIST expression in the officially female sample GSM4697983, we analysed and compared the expression of XIST and several genes located in the male-specific region of the Y chromosome (*Ubiquitously Transcribed Tetratricopeptide Repeat Containing, Y-Linked* (*UTY*), *Ubiquitin Specific Peptidase 9 Y-Linked* (*USP9Y*), *Lysine Demethylase 5D* (*KDM5D*), *Eukaryotic Translation Initiation Factor 1 A Y-Linked* (*EIF1AY*), *DEAD-Box Helicase 3 Y-Linked* (*DDX3Y*), and *Ribosomal Protein S4 Y-Linked 1* (*RPS4Y1*)) [61,62]. For this second analysis (again working for all gene expression data as an independent verification for the male sex), we decided to include all human samples of both studies (Fig. 3). The samples in Fig. 3 are grouped by research group (rows) and labelled according to their sex in the respective metadata. The samples are grouped into control and COVID-19 samples. The colour of the bars indicates the sex (in case the metadata and the sex-specific gene expression correspond, blue indicates a correctly labelled male donor

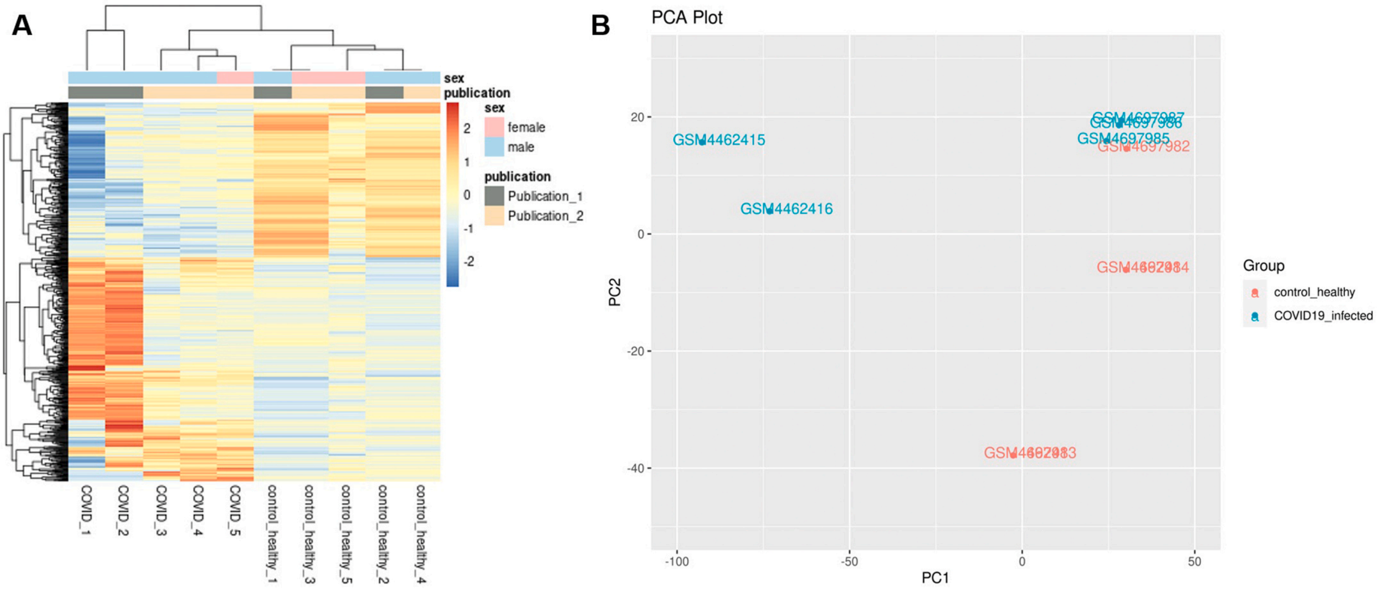


Fig. 2. Heatmap and principal component analysis visualizing the samples of both studies. (A) The similarity between "control_healthy_1" and "control_healthy_3", as well as between "control_healthy_2" and "control_healthy_4" became apparent while visualizing the gene expression in all samples as a heatmap. (B) The PCA indicates differences and similarities between the samples. GSM4462413 and GSM4697983 as well as GSM4462414 GSM4697984, are superimposed. Three of the COVID-19-infected samples of the data generated by Han et al. (2021) [51] also cluster closely but are not superimposed. Red dots symbolize control samples ("healthy" (not COVID-19 infected) according to the metadata), and blue dots denote COVID-19 positive samples (according to the respective metadata). Own figure.

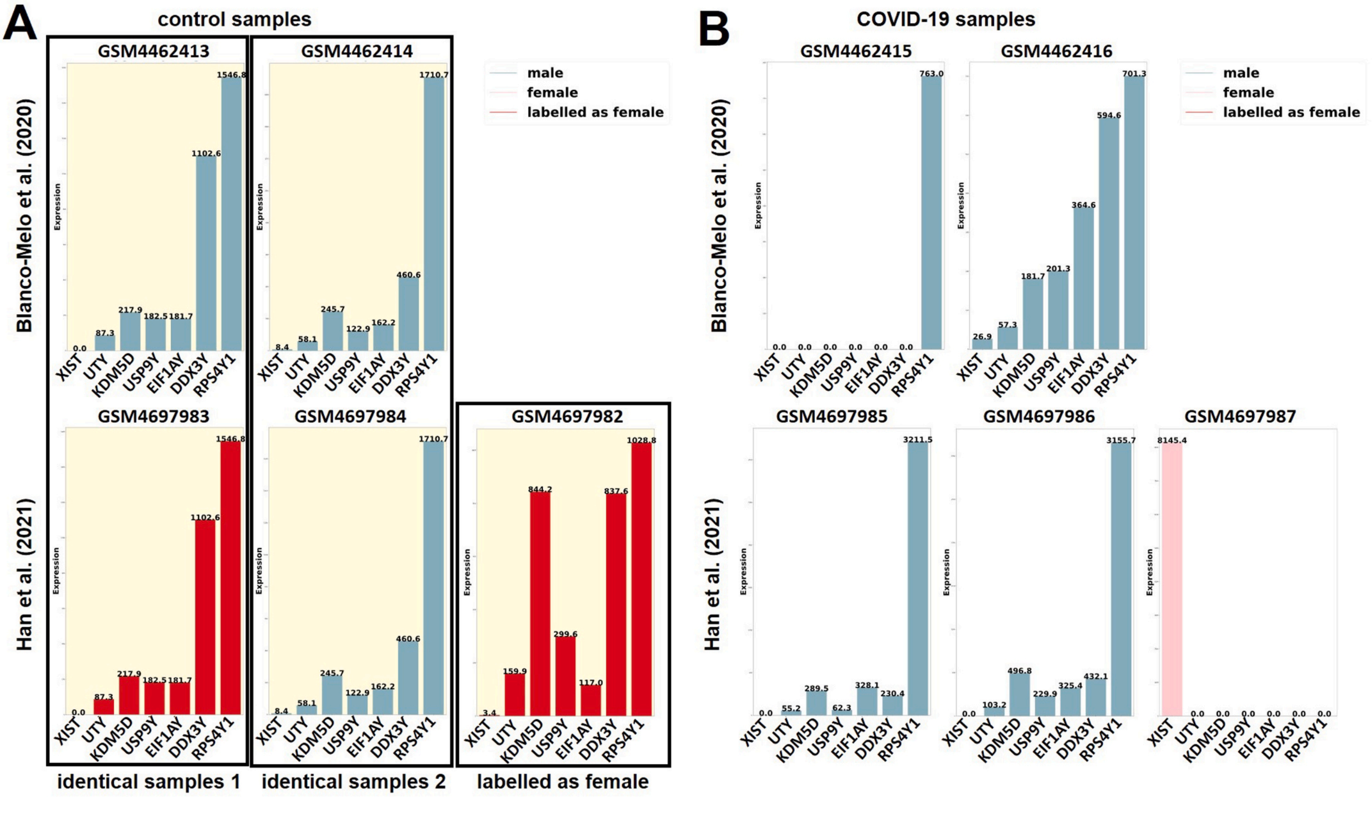


Fig. 3. Sex-specific gene expression in the samples of Blanco-Melo et al. (2020) [50] (starting with GSM446…) and the samples of Han et al. (2021) [51] (starting with GSM469…). (A) Control samples published by Blanco-Melo et al. (2020) [50] and Han et al. (2021) [51]. XIST is usually expressed in females, as it is responsible for the inactivation of the second X chromosome in females [60] (e.g., in GSM4697987, which is a correctly labelled female sample, indicated by the pink bar colour). The other genes are located in the male-specific region of the Y chromosome. Correctly labelled male samples are indicated by blue coloured bars and express Y-specific genes. GSM4697982 and GSM4697983 (that were obtained from female donors according to the metadata), are wrongly labelled as female (indicated by the red colour of the bars and the yellow background). The identical gene expression of GSM4462413 and GSM4697983 and GSM4462414 and GSM4697984, respectively is highlighted by rectangular frames grouping the respective samples together…). (B) COVID-19 samples published by Blanco-Melo et al. (2020) [50] and Han et al. (2021) [51]. Own figure.

and pink a correctly labelled female donor) or an error (red bars if the information regarding the donor's sex does not fit the sex-specific gene expression). Additionally, all samples which appear to be affected by a metadata error are highlighted by a yellow background colour. The similar samples and the sample with the wrongly annotated sex are further emphasised by rectangular frames and labels indicating the error.

The samples starting with GSM469… are part of the dataset by Han et al. (2021) [51], and the samples starting with GSM446… are part of the dataset by Blanco-Melo et al. (2020) [50], the sex is indicated according to the data available via the respective publications and the expression of sex-specific genes. The gene expression of the samples in question (indicated by black frames) is exactly similar for all of the analysed genes, which is highly unlikely if the samples were obtained from different individuals and sequenced using different sequencing platforms. After our inquiry regarding similar samples, the authors demonstrated responsibility and rectified the situation. As of January 5th, 2022, GSM4697983 and GSM4697984 are labelled as reanalysis of GSM4462413 and GSM4462414, respectively, in the GEO database (indicated by black rectangles in Fig. 3). Thus, the striking similarity of the samples and the striking similarity of the raw data was indeed due to the samples in question being the same samples, which is now (since January 5th, 2022) indicated at the individual samples' GEO entries. During our second analysis (which is depicted here), we used all human samples (COVID-19 infected and healthy controls) of both publications, instead of only analysing the similar samples, comparing the groups COVID-19 infected vs. healthy (as indicated in Fig. 2). The bar plots in Fig. 3 show the expression of XIST and the above-mentioned Y-specific genes. The third healthy sample of Han et al. (2021) [51], GSM4697982, which was obtained from a healthy female according to their metadata, shows a significantly lower XIST expression compared to GSM4697987. This sample was obtained from a female COVID-19 patient according to Han et al.'s data [51]. Additionally, GSM4697982 shows the expression of several Y-specific genes. Hence, we contacted the authors again. They responded immediately and asked for further information regarding our analyses. Thus, we provided detailed information regarding the analyses, including the bar plot shown in Fig. 3 and a list of the Y-specific genes. Since there was no further email exchange (September 2023), we assume they are still checking their data. Recognising the authors' diligent approach to the GEO database entries, we are confident that they will address and update the information on the reanalysed samples and the inaccurately annotated sex of their third control sample in their publication.

6.6. Identifying incorrect metadata and remaining doubts such as incorrect labels

Though XIST and Y-chromosome specific gene expression allow reconstructing which sex is correctly labelled in the metadata, extensive analysis of the raw data is often required to reconstruct more complex data such as age or pathology. Hence, we advocate in (i) prevention of metadata errors by checks and input routines, (ii) labelling of incorrect metadata, and if necessary (iii) correction of the incorrect metadata. Additionally, (iv) independently audited metadata might get a special annotation or seal of approval so that researchers can easily recognise the reliable data.

The cited authors took corrective measures and updated two of the GEO database entries (updated on January 5th 2023, no further changes (e.g., in the article) as of September 2023). The example illustrates this point, and very often, one cannot blame somebody if such an error occurs as the necessary information for this got lost long before. However, we now should discuss how often such errors occur and what difficulties await researchers trying to correct these errors.

6.7. Strategies for detecting and reducing metadata errors

Metadata errors happen even more often and should be reduced by comparison routines**.** Is an error in the metadata a rare incident? No, mistakes with metadata happen very often, each time data are stored, links between data and metadata got lost or central metadata (experimental conditions, samples, numbers, clinical features) either not entered or mislabelled. However, we can pinpoint mistakes as the one above only rarely, as this requires close comparisons of the control data sets in unrelated publications – or, to be more general, basic and more refined quality checks on transcriptome data sets and other omics data sets across all public data sets. In the following, we provide a brief overview of such errors to stress the necessity of doing such checks.

6.8. Detecting and reducing nucleotide annotation errors

The notion of abounding mistakes in data and metadata is correct, as one can already see with basic mistakes regarding nucleotide errors [63, 64]:

Park et al. (2021) developed a semi-automated screening tool to detect nucleotide annotation errors, Seek & Blastn. Alarmingly, they found 712 papers with wrongly identified sequences chosen from five literature corpora (two journals (7399 articles published in *Gene* (2007–2018) and 3778 published in *Oncology Reports* (2014–2018)) and three targeted, topic-specific larger corpora using specific keywords (single gene knockdown of 17 specific genes (174 articles across 83 journals), articles related to miR-145 (a total of 163 articles), and articles related to cisplatin or gemcitabine treatment (258 articles) resulting in a total of 11,772 articles) [63]. According to Google Scholar, the 712 problematic articles were cited 17,183 times in March 2021, including clinical trials [63].

At the time of publication, up to 4% of the problematic papers in each corpus had already been cited at least once by clinical trials [63]. However, Park and colleagues also analysed the articles further and predicted a high probability of 15–35% of these problematic publications to be cited in future clinical research, based on the concepts contained within the articles [63]. Hence, there is serious concern that about a quarter of these publications will likely impair clinical research by misinforming or distracting the development of potential cures, especially as the majority of the problematic articles has remained uncorrected [63].

6.9. Detecting and reducing annotation errors regarding sex

Toker et al. (2016) [65] used human transcriptomics studies to compare the sex of the subjects that were annotated in the metadata with what they termed the "gene-sex", the sex of the subject determined by analysing the expression of the female-specific gene XIST and the male-specific genes KDM5D and RPS4Y1 [65]. For their study, they analysed the gene expression of the female specific gene *X-inactive specific transcript* (*XIST*) and the male-specific genes *Ribosomal Protein S4, Y-Linked 1* (*RPS4Y1*) and *Lysine (K)-Specific Demethylase 5D* (*KDM5D*) in 70 human gene expression studies (containing a total of 4160 samples) which had the sample donor's sex annotated [65]. Their analysis revealed that 46% (32 datasets) of the 70 datasets examined in the study contained mislabelled samples, expressing genes they should not be able to express according to the information in the metadata [65].

As 29 of these datasets were associated with a publication, the authors had a closer look at the respective original studies and tried to find out whether the incorrect annotation was solely due to a miscommunication while uploading the data on the GEO database. Twelve of these 29 studies provided enough information in the publication to show, alarmingly, that the discrepant sex labels had already been present in the publication [65]. Thus, in at least 12 studies, the annotation error had already been present in the original publication and was not due to a miscommunication while uploading the data to the GEO database [65].

Finally, Toker and colleagues compared four datasets that used samples from the same collection of subjects. Although not all of the four studies analysed all of the available samples of the collection, Toker et al. (2016) reasoned that if the collection contained incorrect metadata, the subsequent error should affect all of the four studies. However, while two analyses contained mismatched samples, [65] the mismatched samples differed between the two datasets [65]. Additionally, the respective samples were correctly annotated in the other two studies, indicating that the samples had been mislabelled in the respective studies instead of an error while recording the subject's sex [65].

Checking for gender-labelling errors might be relatively easy and cost-effective. This is indicated by a method for predicting gender-labelling errors using X-chromosome SNPs by Qu et al. (2011) [66]. Their method simultaneously accounts for heterozygosity and relative intensity of X-chromosome SNPs in candidate genes and does not require Y-chromosome data and no additional space for gender-prediction SNPs in the genotyping set [66]. Using only nine X-chromosome SNPs in two candidate genes, they were able to predict several sample switches accurately [66]. Additionally, their prediction step requires no additional experiments in the laboratory and can be performed on various different sample types [66].

6.10. Detecting and reducing transcriptome metadata errors

Mishandling of metadata is always possible for transcriptome data sets, as well as for all other omics data sets (proteome, phosphorproteome, metabolome, genomics), and the problem is that wrong metadata, wrong labels, wrong conditions, and mislabelled controls are very difficult to spot and to correct in retrospect. The control is best done using all the information present at submission.

The big wave of incorrect annotation and significant data errors is steadily rising:

The sheer amount of data and its rapid growth further complicates finding such errors.

Around 2013, the amount of publicly accessible gene-expression data sets was about to hit the one million datasets milestone [9]. Using these data was becoming a valid method of gathering research data, as 20% of the data sets deposited in 2005 had been cited by 2010 [9]. Additionally, 17% of the data sets deposited in 2007 had been cited by the end of 2010 [9], underlining the increasing importance of the GEO database. By the end of 2020, GEO entries reached 4 Million, and only a year later, we were at 4,713,471 samples (November 2021), which has by now (September 2023) increased to 6,670,188 samples. The natural increase in data volume without improved metadata reporting quality controls will lead to ever more errors. The impressive number of publications citing the use of the same publicly available dataset (e.g., the comprehensive dataset by Fleischer et al. (2018) [5]) [6], which is steadily rising, further demonstrates the urgent need for correct metadata in every dataset but, as our resources for checking are scarce, with special focus on highly cited master datasets as so much research depends on these key datasets.

6.11. Detecting and reducing gene name autocorrect errors

It has been known since 2004 that some gene names, such as MARCH3, SEPT8 and DEC1, can be autocorrected into dates in spreadsheets [67]. In response to this issue, a 2016 publication by Ziemann et al. raised awareness and prompted the Human Gene Name Consortium (HGNC) to rename genes with names less prone to autocorrect [67,68]. However, almost twenty years after realizing the problem and five years after that article, Abeysooriya et al. (2021) report that despite awareness of the problem and measures taken by both the HGNC and by software developers, the number of Excel files containing gene name errors even increased [67]. In addition to giving tips for preventing such errors, the authors also set up an automated reporting system, which is available at http://ziemann-lab.net/public/gene_name_errors/ [67]. Another approach besides rising awareness and renaming the respective genes would be to consider foregoing the use of Excel and opting for the CSV-format and the use of software without autoformatting. Additionally, due to CSV being a non-proprietary format, this approach would align with Tim Berners-Lee's 5 Star Linked Data Principles [22,23] for Linked (Open) Data. The cumulative criteria include that the data is required to be available on the Web (the first star). In order to qualify as Open Data it also needs to have an open licence for being available on the web. Furthermore, the data should be in a (non-proprietary) machine-readable format (the second star is obtained for a machine readable format, the third if this format is non-proprietary) [22,23]. A fourth star is awarded for the use of RDF and SPARQL (query language to query RDFs), the open standards from the W3C, so that the data can be referenced by others [22,23]. Ideally, the data should also be linked to other data to provide context (resulting in a five star rating) [22,23]. A combination of both approaches, awareness of the potential side effects of autoformat and autocorrect features as well as the use of non-proprietary data formats that are less prone to autoformat and autocorrect, such as CSV, could help preventing such errors. However, researchers creating and reading CSV data still need to be aware of this potential error source to avoid autoformat and autocorrect errors if they prepare or open CSV-files using spreadsheet software applications that offer autoformat and autocorrect options.

6.12. Detecting and reducing citation errors

Digital object identifiers (DOIs) are part of the bibliographic metadata in Crossref, which is provided by the publishers and not double checked by Crossref [69]. Thus, Crossref faces similar challenges as other databases containing metadata that have not been double-checked. Additionally, DOI-mistakes have been analysed reported in other databases such as Scopus, Web of Science and PubMed [69], suggesting that this citation-related metadata problem is widespread across databases and not specific to Crossref. In addition to analysing the taxonomy of the DOI errors, Cioffi et al. (2022) also developed a cleaning mechanism that could be used to correct mistakes in DOIs automatically [69], which gives reason for hoping that their tool and similar approaches might help coping with the flood of data and the concomitant wave of errors.

A 2018 article by Brembs indicates that prestigious journals often appear to struggle achieving equally high reliability regarding data and metadata compared to other journals [70]. Reason include that these journals also receive the highest number of submissions and thus are faced with a monumental task of checking data, metadata and consistency [70]. Moreover, often time is critical, competition fierce and latest methods used are just starting to become reliable charting unknown, new territory. As journals are required to carefully examine each manuscript, including all attached supplementary data and metadata, there is an urgent need for tools that help both scientists and journals to cope with the ever-growing flood of research data. These tools should assist scientists in documenting and archiving their data and allow journals to easily and ultimately automatically double-check and verify this data. Additionally, the use of such tools should allow to update large-scale data including supplements over time as more data become available, marking clearly version history to document. This ensures reproducibility of the reported research as well as confirmatory data gathered only later like is already standard in large-scale database. Thus, ideally, these tools should be standardised, enabling easier and more reliable data review and analysis for all involved parties.

6.13. The battle for improving metadata quality has just started and must not be lost

6.13.1. Hope for transcriptome metadata: comparison and consistency checks

Our review and examples of transcriptome data and errors show that metadata must be rechecked carefully. Moreover, as a general rule, also the repositories should have automatic checking routines for such mistakes. This is easy to achieve, at least for transcriptome data: if every entry is checked for the novelty of the raw data or even the partial overlap with the stored data, it is easy to identify this type of mistake. Sharing omics data such as RNA-sequencing data as publicly available datasets offers excellent value to the research community, as computational analysis of already existing data can save time and resources. Additionally, analysing already available datasets with other methods, new tools, or a different focus can generate new insights and is slowly becoming standard practice and can even reveal further insights [6].

Based on the studies above, it is reasonable to assume that individual metadata errors are normally distributed and that a fatal binary error, such as wrong sex or mix-up of control and treatment, occurs for a low percentage of publications (at least 1–3%). At the same time, a substantially larger fraction (roughly estimated about 5x more) has minor quality issues.

6.13.2. The importance of spotting metadata errors in all data types

Spotting metadata errors in all data types requires effort but is essential: Unfortunately, not all labelling errors can be found as easily as the wrong sex in transcriptome metadata. Genetic data are in principle comparable by similar techniques (most easily by mapping genomes against each other), but genetic variation is the key in sequencing new DNA, and hence, wrong labelling of sample and treatment and specific sample conditions may go unnoticed as the variation caused by this is hidden in the “expected” “natural” variation. Even with machine-learning techniques, detecting metadata errors can be quite challenging due to the presence of natural variation.

Therefore, machine-learning models may struggle to differentiate between true errors and inherent variability, leading to unsatisfactory predictions [71].

Most errors regarding the metadata of other omics data types are complicated to spot in retrospect. For instance, exact conditions in proteomics experiments or time points or conditions, sample preparation and handling are challenging. Even more challenging are metabolomics samples, as sophisticated techniques are used, different protocols are available, and sometimes critical information to allow cross-comparisons over different datasets is not available, rendering cross-comparisons impossible. Samples that erroneously got labelled as control samples are harder to identify and might cause even more damage to research, especially if only a limited number of samples with the particular condition are available.

Finally, metadata in imaging data are comparatively easy to spot if the error pertains to the annotation and what is visible on the image. This should improve the more powerful computer-assisted annotation or even automatic annotation becomes. Tools such as the MetaData Editor for microscopy MDEmic, which allow editing and creating of detailed metadata can improve the data interoperability of imaging data [72] and thus help to apply the FAIR (findability, accessibility, interoperability and reusability) principles [19] in research [72]. However, all metadata errors not directly visible from the image, such as sample preparation, harvesting conditions, pharmacological treatment and time points, are again difficult to spot and require extensive cross-comparisons. For other, more functional data and the typical “individual molecular biology experiment”, the same considerations apply. Thus, we might have a reason for cautious optimism as long as more cross-data checks are purposely or even systematically applied.

7. Coping with the Big Data Wave

A possible approach for coping with the Big Data Wave is by keeping metadata quality high and making systematic comparisons. Big data are constantly increasing, which inevitably increases the workload on the people handling them. Unless there are automatic quality controls, cross-comparisons to validate metadata, and checks and counter-checks ensuring that the entered information is correct, we are to drown in errors as the number of personnel involved in databanks is undoubtedly not increasing at the same pace as data generation.

The well-known reproducibility crisis [73] is triggered by a difficult-to-avoid bias in publishing positive results, not showing negative results, or even omitting most of the confounding data. There are the correct reservations of statisticians regarding statistical biases, too small samples, extravagant claims and missing controls [74,75]. However, from the start of any scientific study, good data need good curation. If there is no exponential increase in automatic data and metadata quality controls, we will experience a steady decline in data quality, inversely proportional to data growth.

8. Strategies and obstacles on the road to data integrity and reusability

To be valid, data need to be correct, complete, readily available, and accessible, and compatible; otherwise, the data will not easily be used [76]. This also includes the comparability of data, which is necessary for analysing and comparing multiple datasets created by different researchers. A substantial heterogeneity, as reported by Perumal and colleagues regarding the data quality in anthropometric studies [77], can impair studies based on several datasets and limit research.

Several concepts have been created and continuously developed to meet these requirements. An international data quality (DQ) research collaboration developed a harmonised intrinsic data quality framework (HIDQF) with two contexts for DQ assessment (verification and validation) and three DQ categories (conformance, completeness and plausibility) [78] to assess the intrinsic quality of a dataset. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines have been established to evaluate the data quality in research publications and are continuously being developed [79]. These guidelines serve as a possible resource to assess the data quality in the published literature [78].

Additionally, the need for a sufficient infrastructure allowing the reuse of scientific data has led to the design of the FAIR (Findability, Accessibility, Interoperability, and Reusability) Data Principles, which take both human researchers and so-called “computational stakeholders” (computational agents and applications for data retrieval and analysis) into account [19]. As Wilkinson et al. (2016) emphasised, FAIRness is vital for proper data management, which is a precondition for other researchers’ reuse of scientific data [19].

In addition to these guidelines and good practices, several computational methods are being implemented to solve the data dilemma. Various DQ assessment tools, both commercial and open-source [78], are available and often shared by the developers, e.g., via GitLab [80]. However, Liaw et al. (2021) observed that only a few studies reported their datasets’ quality [78].

In clinical practice, informatics is being used to tackle a similar problem: Reporting and analysing patient safety events (PSE) and the measures taken to prevent them are often impeded by missing or incomplete data [81]. In their assessment of narrative medication error reports, Yao and colleagues observed that the narrative parts of PSE reports contained extensive and valuable information. At the same time, structured fields were often ignored [81]. A possible solution for this dilemma is the use of natural language processing (NLP) tools since a proof-of-concept study has already demonstrated that existing NLP systems, such as the Averbis Health Discovery tool, can extract medication information from narrative texts, e.g., from unstructured medical discharge letters [82]. Tools like that could be able to compare the text of the publication to the metadata and help annotating the metadata correctly or finding discrepancies between metadata and publication.

Although NLP tools show promising results in automated text extraction [82], solely relying on machine learning might add potential uncertainty. Additionally, a recent evaluation of the large language models (LLMs) (“chat”) GPT-3.5 and GPT-4.0 has shown that their performance and behaviour can change substantially [83]. Chen et al. (2023) evaluated GPT-versions and report that some tasks were solved substantially worse after a relatively short amount of time (between March and June 2023), which indicates a need to continuously monitor the behaviour of LLMs [83].

Furthermore, before monitoring the long-time behaviour of an LLM or other ML approach can even take place, the performance of new models has to be evaluated and probably also compared to already existing models performing the same or a similar task. Objectively comparing different ML models can be challenging as different studies might use different metrics [84]. Moreover, different ML models for different uses might have different requirements, which can in turn affect the evaluation of a tool [84]. An example is the balance between precision and recall, the two most used measures for evaluating the performance of applications for pattern recognition and information retrieval [46].

Precision is defined as the relation of the number of correct results (True Positives, overlap of the two circles in Fig. 4) and the number of all results [46]. In a database search, this would equal the number of relevant documents that were retrieved divided by the total number of documents that were retrieved [47]. Recall is defined as the relation of the number of correct results and the number of expected results [46]. In a database search, this would equal the total number of relevant documents that were retrieved divided by the total number of relevant documents [47]. Data that fulfils the criteria of findability and interoperability is persistently identifiable and re-findable, machine-actionable and its metadata is ideally syntactically parseable as well as semantically machine-accessible [37]. This will result in high recall and high precision, meaning that all (or almost all) relevant data have been found, and all (or almost all) of these data were correctly identified as relevant.

Maximising recall is related to a low number of false negatives and assigning more instances as positive, which increases the number of "false alarms" [84]. At the same time, a high precision requires a low rate of false positives, which can result in missing some positive events as only very strong positive predictions will be returned [84]. While in cancer detection some false alarms are tolerable, a less severe and more prevalent disease might require a higher precision [84]. Moreover, a combination of low precision and high recall and the resulting false alarms will unnecessarily increase the manual workload and waste time [84]. Incomplete metrics can also lead to confusion or even give a false impression of the model's performance, as Hicks et al. elegantly elaborate in their 2022 publication in *Scientific Reports* [84].

In database searches, precision and recall are also of interest: Researchers investigating the impact of a specific disease on gene expression in a particular tissue type should be able to effortlessly find (almost) all relevant datasets for their analyses. For instance, datasets that include omics data from both healthy and disease-affected samples derived from the specific tissue type. While precision and recall are used to evaluate the accuracy and relevance of the retrieved information, the afore mentioned aspect of findability primarily concerns the discoverability of data, which can be enhanced by adding sufficient metadata.

8.1. Domain-specific simple annotation tools

These can be a pragmatic solution and cover diverse areas:

Bacterial genomes: Since the verification of supporting data and identification of errors and inconsistencies are challenging tasks, Schmedes et al. (2015), have developed an automated, easy-to-use Excel-based tool for the curation of local bacterial genome databases, which can be used as a quality check before downstream analyses are performed [85]. Additionally, they also emphasise the importance and the urgent need for additional tools and quality control practices, and suggest that an upfront quality control of data by public database managers would save downstream resources and provide the end user with better quality data and metadata [85].

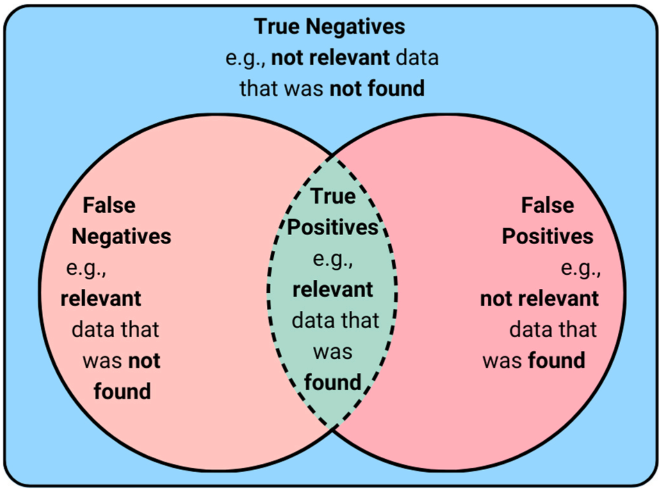


Fig. 4. Visualisation of True Positives, True Negatives, False Positives and False Negatives. Own figure.

Sequence read archives: Crandall et al. (2023) report missing spatial and temporal metadata in genome-scale genetic diversity data, which hinders the reuse of these data for monitoring programs and other purposes. They report that in 2021, only about 13% of the over 300,000 Samples in the Sequence Read Archive (SRA) that might be relevant to global genetic biodiversity contained information about the precise location and time where they were obtained [86]. Additionally, they observed a rapid decrease in the availability of metadata necessary to restore the missing information [86]. Due to the rapidly declining metadata availability, which they found mirrored in other kinds of biological data, they raise attention for the need for updated data-sharing policies and researcher practices, as metadata contain valuable context which should not be lost to science forever [86]. Besides this (potential) data loss, the absence of appropriate spatiotemporal metadata additionally represents a loss of research effort, which could range from tens to hundreds of millions of dollars and also affects the Indigenous peoples who otherwise could possibly have benefitted from genomic information originating within their territories [86].

To tackle this problem, the group (12 professional researchers and 13 graduate students) spend about 2300 h trying to restore the missing information during their "datathon" (data restoration competition), which is described in detail in Crandall et al.'s 2021 publication [86]. Their effort to retrieve the missing metadata could rescue over US $2.1 million worth of genomic sequence data [86], which indicates that trying to restore and correct missing metadata is worth the effort. However, this is not an ideal long-term solution, it would be much better if metadata were shared as diligently as primary data are already shared because only the added metadata make primary data FAIR [86]. Hence, the authors also provide a detailed list of required and recommended metadata that might be of interest for monitoring genetic diversity, including definitions of the terms [86].

Sequence database cross-check: Missing (meta)data is not the only challenge for bioinformatics. Increasing evidence suggests that sequence databases harbour significant amounts of erroneous information, including spelling errors in protein descriptions, contamination of sequence data, duplication, and inaccurate annotation of protein function [87]. Therefore, Goudey et al. (2022) analyse and describe the interconnectedness and interdependency of different databases and how the relationships between records in these databases can be employed to understand and improve the quality of sequence records and data in sequence data bases [87]. They propose regarding the various sequence databases as parts of a greater whole, instead of seeing them as independent entities that are loosely linked [87]. This *sequence database network* and the relationships between records and machine-learning methods, such as trust propagation techniques, can be exploited to detect and correct annotation errors as well as for verification of connected records [87]. Additionally, they highlight the need for new metrics for quantifying quality of records and their respective metadata and the importance of propagating updates or corrections [87]. Machine-learning models, such as random forests and artificial neural networks, can further be employed to find sequencing errors [71]. Another important aspect of using big data is data cleaning or data cleansing, which aims to improve data quality via the identification and the subsequent removal of errors [88]. “Dirty data”, which is defined as being inaccurate, inconsistent and incomplete data, and poor-quality data can affect analyses [88]. Data scientists spend a great amount of time and effort on cleaning and organizing data, taking up 50–80% of their work time [89]. While some errors are relatively easy to spot, such as missing values that get encoded as an unrealistic number (e.g., 99 years of education due to missing values being encoded as 99) [89], other errors, such as the error we happened to find by accident are more difficult to identify. Since identical data was labelled with different metadata, the only method to identify this error would have been to compare the newly uploaded sequences to all other sequences in the database. By doing so, the duplicate sequences would have been identified, and a subsequent comparison of the respective metadata would have indicated errors in the metadata. This approach would require a huge amount of computing power and is therefore impractical. Routinely comparing the annotated sex and the sex-specific gene expression can indicate errors in the metadata, but only the researchers who created the data might be able to find out whether the respective samples have a simple error in their metadata or if the sample in question got tagged with the metadata of another sample. This demonstrates the huge responsibility for researchers sharing their data.

Enhancing metadata from lab experiments: A possible method to handle newly generated laboratory data and the relevant metadata, such as who created the data, using which samples, following which protocols, has been described by Panse et al. (2022), who explain how the life sciences community of the Functional Genomics Center Zurich (FGCZ) is able to “glue together” data, including metadata, computing infrastructures, such as clouds and clusters, and visualisation software using their self-developed B-Fabric system, allowing instant data exploration and ad-hoc visualisations [90]. They also present their lessons learned, which is not only valuable information for researchers facing similar data organisation tasks but also showcases the qualities and advantages of their software solution [90].

Ontology and terminology corrections: Another important aspect is highlighted by Beretta et al. (2021): They bring to attention that interdisciplinary data sharing and the discovery and reuse of data face additional challenges due to discipline specific formats and different metadata standards as well as semantic heterogeneity [91]. Therefore, they introduce a user-centric and flexible metadata model, which is based on a common paradigm based the observation concept [91]. In accordance with the FAIR principles, they aim to reuse existing ontological and terminological resources and specify the semantics of the elements of the model with ontologies and vocabularies [91]. By adding semantics to metadata, the model enhances discoverability and semantic interoperability, which enables interdisciplinary research projects [91]. Additionally, the model can utilize reasoning capabilities and for instance enhance a search query for a certain fish in the ocean by finding all datasets of the various species of the fish as well as datasets related to the habitat of the fish [91]. There are of course more solutions, for instance the tool Protégé (https://protege.stanford.edu) is an ontology editor equipped with the capability to be integrated with a reasoner for ontology consistency testing (e.g. HermiT).

Preserving metadata for heterogenous data repositories: Discovering, querying and integrating challenging heterogenous data and knowledge from different sources, has been analysed by Kamdar and Musen (2021) [92]. They meta-analysed more than 80 biomedical linked open data sources within the *Life Sciences Linked Open Data* (LSLOD) cloud, which has been created using Semantic Web and linked data technologies [92].

In linked data, information is linked from different sources via a site [21]. This approach can even connect different databases which are maintained by separate organisations or heterogeneous systems which did not easily interoperate at the data level [21]. By using a Life Sciences Linked Open Data (LSLOD) schema graph, [92] observed that there is still need for improvement, as the LSLOD cloud was not as densely connected as assumed and that several databases were not well connected to others or even not interconnected with other databases at all [92]. This demonstrates the adverse effect of the heterogeneity and the quality discrepancies of the LSLOD sources and the lack of common vocabularies [92] and highlights the need for transforming non-FAIR data into linkable data [21] as well as the importance of the Linked Data Principles. These principles have been described in detail by Tim Berners-Lee, who highlighted and explained the importance of URIs and the information that can be provided via URIs. Additionally, he introduced a five star rating scheme, 5 Star Linked Data, to point out the important aspects of Linked (Open) Data [22,23], which have been described above. These principles can also be adapted for specific requirements, for instance, to fulfil the requirements of the Linked Open Data Cloud, which include additional requirements including a resolvable http:// or https:// address, and being connected to data that is already part of the diagram via RDF links [93]. These strict criteria might also explain missing entries, since the respective entries might have lacked some of the required criteria.

Another approach besides employing Semantic Web techniques, is decreasing the semantic heterogeneity, e.g., by increasing vocabulary reuse. Using a defined vocabulary will enhance the effectiveness of querying and enable integrating diverse biomedical sources in the future [92].

This might be aided by different tools that function similar to the tool METAGENOTE (referring to the collection of METAdata of Genomics studies on a web-based NOTEbook), which has been developed to help researchers using standardised metadata describing their genomics samples during the submission to the Sequence Read Archive [27,94].

Github supported open tools for repositories and data maintenance: Ideally, these tools for checking the data should be available as open-source software. For various analysis tools, this is already common practice and many developers also maintain GitHub repositories, which offers the opportunity to interact with the developers and other users. This allows users to draw attention to problems and ask for support when needed. Additionally, all questions, discussions and solutions are archived and accessible for future reference. How well this system works can be seen in various GitHub Repositories for open source R tools such as DESeq2 [57], Seurat [95–99] or OmniPath [100,101].

This practice is not only convenient but often also a requirement upon publication of an article introducing the tool, and should be implemented for publications introducing new datasets and/or using these analysis tools as well.

An example is the *Nature* publication “*A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types*” by Kozareva et al. (2021) [102]. The authors made their data available via the GEO database [8] and the Single Cell Portal, which is available at https://singlecell.broadinstitute.org/single_cell, and additionally created a GitHub Repository (available at https://github.com/MacoskoLab/cerebellum-atlas-analysis) where they describe in detail how others can recreate their figures [102].

Another possible approach for addressing deficiencies in data quality or the absence of corrections might be community moderation, which could give users the opportunity to discuss articles or research findings or even highlight discrepancies. However, solely relying on community input without moderating options or even an “anti-censorship” philosophy, such as the late 1970 s online Bulletin Board System CommuniTree [103], is a risky strategy. The example of CommuniTree, which has been described in detail by Seering (2020) demonstrated that the dream of the internet being a “market place of ideas” which would allow better perspectives to naturally rise to the top, was a utopia [103]. While the consequences of a more heterogenous user group might have been a surprise for the first CommuniTree users, today, online conflicts are a well-known and unresolved problem [103]. Therefore, some forms of moderation options need to be implemented. The moderation itself can either be performed by the platform itself or by the community, for instance by volunteer moderators [103]. However, both approaches require additional resources, either by the respective platform itself or by volunteers or even by volunteers and the respective platform, since the platforms might want to employ platform administrators who are in charge of final content moderation decisions [103]. An example for the usefulness of an option to comment scientific articles is the *bioRxiv* platform, the preprint server for biology. The platform links preprints with discussions about the respective articles in the media and in Tweets and even provides links to online discussions regarding the preprints that occur elsewhere (Community Reviews). Additionally, *bioRxiv* offers a Comment option, where readers can discuss the article or ask questions regarding the article, which are sometimes answered by the authors or other readers. An example is the *bioRxiv* preprint of the *Cell* publication by Blanco-Melo et al. (2020) [50,104], which can be found at https://www.biorxiv.org/content/10.1101/2020.03.24.004655v1#comments. Additionally, published preprints, such as the preprint by Blanco-Melo et al. (2020) [104], are linked to the final version of the article [50] upon publication.

8.1.1. *Tools for data management and version control*

Besides the above-mentioned tools and strategies, several software solutions to enhance data management and facilitate generating FAIR data are available. A selection of these tools is presented in Table 4, a more detailed description of these tools including advantages, disadvantages and FAIRness-rating as well as the respective links can be found in the Supplementary Data.

8.1.2. *Selected tools for data handling, error detections, bioinformatic analyses and publications*

Additionally, several tools have been designed and developed to aid researchers in every step of the publication process, from citation management software to workflow management, quality control and error detection. A small selection of these tools is summarised in Table 5, more detailed information, including descriptions, links and last update, is available in the Supplementary Data.

Notable examples regarding big data and cloud use are for instance the open-source projects Apache Hadoop, Apache Spark, and Databricks.

Apache Hadoop, which was developed by Doug Cutting and Mike Cafarella, is one of the well-known solutions for working with big data [105]. Five characteristics, often referred to as the 5 Vs of Big Data, describe big data: Volume, velocity, value, veracity, and variety [105]. Volume is the most obvious and most immediate challenge of big data, as the data does not only need to be stored but also analysed [105]. Furthermore, big data is created with ever increasing speed (velocity) and refers to a variety of data, which can be structured or unstructured, which affects the storage and the analysis of the data [105]. The potential value is the most important aspect of big data [105]. Although the potential value is huge, Big Data needs to be analysed and turned into value [105]. However, the quality of the data can vary greatly. Due to high volume, velocity and variety, not all of the data can be 100% correct [105]. Therefore, the accuracy of the data analysis depends on the source data's veracity [105], which can be defined as "truthfulness, accuracy or precision, correctness" of the data [106].

Hadoop was originally started as a part of the scalable open-source web crawler project Apache Nutch but soon emerged as an independent and top-level Apache project [107]. The first version of Hadoop consisted of the Hadoop Distributed File System (HDFS), an abstraction layer which is responsible for data storage, and the distributed programming paradigm MapReduce, which was used for managing job resources, as the two main components [107]. When Doug Cutting joined Yahoo! in 2006, a dedicated team for developing the project was created, and the tool, which was named after a yellow stuffed elephant Doug Cutting's child used to play with, has been used extensively by the company [107]. By 2009, Yahoo! could sort 1 TB of data in 62 s by using a Hadoop cluster to index its search engine, and by 2010 an ecosystem of tools, such as Hive, Apache HBase and Pig, was developed around Hadoop [107]. As long as the number of functioning computers in the cluster is sufficient, the relatively fault-tolerant Hadoop MapReduce is able to handle hardware failures well and offers a cost-effective way for processing large amounts of data [108], which is one of the reasons why it is still being used, even almost two decades after its inception. During the COVID-19 pandemic, Apache Hadoop and its MapReduce were proposed as an inexpensive and flexible processing and analysis solution for big data processing during the unprecedented data analysis challenge, which arose from the extraordinary and trailblazing sharing of COVID-19-related data [108]. In 2012, Hadoop 2 and Yet Another Resource Negotiator (YARN) were introduced, which allowed the use of Apache Spark as processing engine as well as other processing models by separating the resource management function of Hadoop from the processing layer [107]. While the big data analysis platform Apache Spark is commonly used on powerful computer clusters, Andrešić et al. (2017) could demonstrate that even a single standard computer is sufficient for data analysis with Apache Spark [109]. Using a standard computer with 8 GB RAM and Apache Spark in single-cluster mode, they could confirm that their approach of combining self-organising map software libraries and Apache Spark was still efficient and fast enough, demonstrating that Spark can also be employed by researchers having limited resources [109].

Additionally, Spark is also suitable for cloud computing and is part of the Databricks Lakehouse Platform, where a Spark compute layer is used for querying, processing, and transforming the data stored in the storage layer, decoupling Cloud storage and Cloud computing [110]. The Databricks Lakehouse Platform is compatible with Microsoft Azure, AWS and Google Cloud, and an example for its use in data analysis is the phishing detection tool by [111], which uses a combination of Microsoft Azure, a spark cluster and Azure Databricks [111].

9. Rethinking the correction of errors in scientific publications

Both avoiding errors and finding errors are essential for data integrity. However, it is equally important to correct these errors once they have been found. Metadata and data errors are something natural and happen. They become more and more as the data accumulate, we can only do our best to lower a priori error probability per dataset but can never achieve a probability of zero. As seen in the rising number of

Table 4
Selected software solutions for data management and version control, detailed descriptions are available in the Supplementary Data.

Category	Tools
Artificial Intelligence in Lab and Data Management	Benchling, Biovia, DataRobot, Eagle Genomics, Elucidata, Genemod, Labguru, SciNote
LIMS (Laboratory Information Management Systems)	eLabFTW, eLabNext, LabArchives, LabCollector, Labfolder, LabKey Server, Labstep, LabVantage, LabWare, LIMS, Quartzy, Rspace
Medical Laboratory Datamanagement	ApolloLIMS, ClinCapture, LIMSABC, Medidata Rave, Medrio, SMART-TRIAL
Metadata and Data Management	Arvados, Asana, BioData Catalyst, CKAN, Figshare, ISAtools, Jira, Mendeley Data, OMERO, Open Science Framework, OpenRefine, ProteomeXchange, SEEK, Terra.bio, Zenodo
Version Control and Collaboration	Bitbucket, GitHub, GitLab, Mercurial, Perforce, Subversion (SVN)
Cloud Services	AWS, Google Cloud, Microsoft Azure, IBM Cloud, Oracle Cloud

Table 5
Selected software solutions for data handling, error detections, bioinformatic analyses and publications [a].

Category	Tools
Citation Managers	Zotero, EndNote, Mendeley, Paperpile, Citavi, JabRef, ReadCube, Papers, RefWorks, Cite This For Me
Data Cleaning	OpenRefine, Talend Data Quality, KNIME
Data Repositories	NCBI Sequence Read Archive (SRA), European Nucleotide Archive (ENA), DNA Data Bank of Japan (DDBJ), Sequence Read Archive in the Cloud (SRA in the Cloud), GenBank
Genome Visualization	IGV, MAKER, PANTHER, HPIDB, MITOS
Annotation Tools	GenSAS, Apollo, PubTator
Annotation Databases	BioMart, Ensembl, Ensembl Genomes, NCBI Gene, GeneCards, UniProt, RefSeq, PDB, InterPro, dbSNP, COSMIC
Sequence Alignment and Annotation	BLAST, MAFFT, Prokka, ANNOVAR, Artemis, EMBOSS Transeq, SnapGene Viewer, GATK, CLC Genomics Workbench, SeqMan Ngen
Sex Determination from Genomic Data	FindZX, Sex.DetERRmine, Peddy, PLINK,
Gene Name Errors	GeneNameErrors2020, Gene Updater
Bacterial genomes	Bacterial and Viral Bioinformatics Resource Center (BV-BRC), NCBI Bacterial Genomes, MicroScope, BacDive, JGI IMG, Ensembl Bacteria, BioCyc
Sequence Analysis	NCBI's Conserved Domain Search, InterPro, UCSC Genome Browser, Bioconductor, EMBL-EBI's Search and Sequence Analysis Tools
Microarray Data Analysis	GEOquery, ArrayExpress
Ontology Tools	OntoCheck, Protégé (which is compatible with the ontology reasoner HermiT), SNOMED CT Browser, UMLS Terminology Services, BioPortal, AberOWL, Ontobee, Ontology Lookup Service (OLS), BioBert
Quality Assurance of Data	FastQC, MultiQC, Trim Galore!, Picard Tools, SAMtools, Cell Ranger, Seurat, SCANPY
Workflow Management	Galaxy, Snakemake, Nextflow, Toil, Taverna, Terra.bio, Apache Hadoop

[a] more details in supplemental excel Table 2

retractions, more publications and better technical means lead to more reasons for retractions being found [112]. Due to inconsistencies of how different journals handle retractions, the Committee on Publication Ethics (COPE) published retraction guidelines in 2009, and the 2010-founded blog "Retraction Watch" covered over 200 retractions and logged more than a million page views in its first year of existence [112].

Nevertheless, correcting errors in scientific publications is challenging due to the current public stigma connected to post-publication updates or corrections, especially when such updates are mistakenly perceived as punitive measures or confessions of wrongdoing [113].

Additionally, the self-correction process and the correction of mistakes face considerable obstacles, which were highlighted by Allison et al. in their 2016 *Nature* article [114]. They identified several challenges associated with the correction process, including disincentives against correction (e.g., fees imposed on authors who request the withdrawal of their publication) and barriers, such as journals requiring publication fees for articles bringing attention to previously published works within the same journal [114]. Furthermore, addressing errors officially in a timely manner may be challenging, as editors may be unable or hesitant to take swift action [114]. This can also be attributed to the conflicting priorities of ensuring fairness to the authors during an ongoing investigation and the need to preserve the integrity of the literature [113].

Luckily, large data infrastructure projects improve data governance such as the NFDI (the German National Research Data Infrastructure) initiative, which was implemented by the German Research Foundation (DFG) and fosters research data management (RDM) [115] and several consortia, including the NFDI4Microbiota (https://nfdi4microbiota.de). International efforts become increasingly aware of this escalating problem on data and metadata integrity, such as Elixir (https://elixir-europe.org), the European life sciences infrastructure bringing together EMBL-EBI and more than 220 institutes within 22 countries [116]. The goal of achieving reproducible results requires ever-new solutions for scientific data management, taking advantage of the willingness of the scientific community to achieve the highest data standards and overcoming the existing barriers by a systematic development of standards, tools, and infrastructure, the provision of training, education, and support, as well as additional resources for research data management (RDM) [115,116].

In light of these considerations, we would like to appreciate the authors for their prompt response in rectifying the issue within the GEO database, as it is not always a given that such corrections are made swiftly. In this sense, our chosen introductory example is a best-practice example. However, it is important to note that the metadata within the article remains inaccurate (as of September 2023), which might be misleading for researchers only considering the information in the article, the supplementary data, and the metadata provided via the SRA Run Selector, without reading all GEO entries for the samples they have chosen to repurpose. At the same time, we would like to draw attention to an optimal practice of considering every available piece of information, even though this is a time-consuming step, especially for big data sets. In this digital age, various technical solutions can address these inaccuracies, for instance by applying the concept of "living articles" described by Barbour et al. in 2017 [113]. Ideally, this leads to a transparent and comprehensive history of changes, which is – in accordance to the key principles for amendments postulated by Barbour and colleagues – accessible for both human and machine readers [113]. Embracing these principles and adopting approaches like the "amendment" system proposed by Barbour et al. (2017), which uses a more neutral term for describing post-publication changes [113], would contribute to a more robust and accurate scientific literature and enhanced reusability of research data. As resources are always scarce and the data avalanche is constantly increasing, at least the highly cited key datasets should be systematically supported by such a regularly updated curation history. Flagship databanks such as EMBL database and GenBank are good examples of continuous curation with new bimonthly releases and daily updates [117].

Another aspect of big data to keep in mind is that every online resource, every tool, and every workflow using these tools is vulnerable to updates as updates might affect their functionality. Thus, if they want to guarantee the usability of their tools, the developers have to keep checking and updating their tools or software packages. For workflows describing or documenting how the results of a publication were generated, this might not be feasible, thus, in these cases documenting the software versions is of utmost importance. If users encounter a workflow that is not functionable, they can check the software versions and may try to recreate the workflow by installing older versions of the software or adapting the workflow to the current software requirements.

9.1. Review limitations and implications

This is not a systematic review of data, metadata and the current best practices for data governance, nor a systematic review about all existing errors in current databases. We want to raise awareness for errors and mistakes in metadata annotation and point out typical errors and helpful metadata maintenance tools. Metadata errors may arise due to problems in sample tracking, and might be avoidable by using appropriate laboratory information management systems and thoroughly documenting the sample metadata. There are errors which could have been avoided as they might be attributed to factors such as a too complicated or too time-consuming procedure for uploading metadata or maybe a lack of awareness regarding how to generate easily reusable data and what needs to be considered. However, most of the errors in metadata just happen with non-zero probability and in this sense cannot be avoided and the data can easily deteriorate with exponential increasing data volume without any counter measures. We hope that this review brings to attention that in big data even little mistakes might end up having huge consequences, especially in biomedical research, and that the errors that have been found might as well be only the tip of the iceberg.

While our review acknowledges necessary scope limitations and the vastness of the topic, it is clear that more studies and systematic evaluations are needed to fully grasp the extent of the issues at hand. One key aspect that emerged from our analysis is the necessity for individuals to possess fundamental knowledge and a sense of responsibility when dealing with data. With big data comes big responsibility, and it is crucial that we foster a culture that promotes accountability and encourages the timely identification and rectification of errors. Thus, to encourage this, our perception of errors in scientific research needs to evolve towards a mindset that acknowledges the occurrence of errors as normal instead of catastrophic events, and focuses on rectifying them promptly and effectively. By removing the stigma associated with mistakes, we create an environment that encourages open communication, timely error identification, and effective remediation, enabling continuous improvement. Overall, the spirit of science lies in the exchange and sharing of information to expand our collective knowledge. Digital data sets in bioinformatics are prime examples of resources that can be easily shared, and they provide opportunities for diverse perspectives and fresh insights.

By embracing responsible practices, encouraging data sharing, and fostering a supportive scientific community, we can navigate the challenges posed by big data and metadata management, paving the way for reliable research. We hope that we could raise awareness for the importance of metadata for future research and a little overview about the efforts that are being made to help avoiding such mistakes in the future.

10. Conclusion

Big data and the accompanying metadata create both chances and challenges for scientific research. The exponentially growing amount of data and the possibly drastic, indirect consequences of mismanaging metadata, akin to the hidden depths of an iceberg, emphasize the need for a comprehensive understanding of the importance of data integrity and a responsible maintenance approach. Automatic consistency checks on metadata integrity should be further improved (sex, age, experimental conditions) and be generally applied. Three-month updates and error corrections are routine in large public databases, but this should include all published large-scale datasets, praising authors for such efforts and not blaming them. Data and metadata integrity are a continuous effort for all scientists – and actually a battle for data quality we must not lose.

Author contributions

AC did make all data processing and data comparisons, supervised by TD. AC, SD, and TD analysed the resulting processed data and data comparisons together. AC drafted the paper; AC, SD, and TD edited and polished the manuscript. All authors agree to the final version. Contributor roles: AC: conceptualisation, data curation, formal analysis, investigation, methodology, visualisation, writing – original draft; writing – review and editing; SD: conceptualisation, formal analysis, investigation, validation, writing – review and editing; TD: conceptualisation, formal analysis, investigation, supervision, writing – review and editing.

Funding

We thankfully acknowledge funding by Stanford Translational Research and Applied Medicine (TRAM) grant, 2020–21 (SD), Stanford Diabetes Research Centre Pilot and Feasibility Grant, 2021 (SD). We are indebted to Deutsche Forschungsgemeinschaft (DFG), grant 492620490 – SFB 1583 /INF (TD, AC) and Land Bavaria (contribution to DFG grant 324392634 - TRR 221/INF) (TD) for their constant support of the respective data infrastructure which stimulated this article. The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors state clearly that they did not use any generative artificial intelligence (AI) and AI-assisted technologies in the writing process.

Conflict of interest statement

All authors declare that they have no conflict of interest. There are no financial, personal or other conflicts of interest by any of the authors (TD, AC, SD).

Data Availability

All data of this manuscript are contained in the paper and its supplementary material. Moreover, the case study example used for illustration has already been deposited in the Bio archives, BIORXIV-2021–473021v1-Dandekar.pdf. The discussed RNA-Sequencing data was made publicly available by the respective research groups and can be accessed via the GEO database (GSE147507 (online available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147507) for the data published by Blanco-Melo et al. (2020) 50 and GSE155241 (online available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155241) for the data published by Han et al. (2021) [51]).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.10.006.

References

[1] Beckett AH, Cook KF, Robson SC. A pandemic in the age of next-generation sequencing. Biochemist 2021;43:10–5. https://doi.org/10.1042/bio_2021_187.

[2] Maher B, Van Noorden R. How the COVID pandemic is changing global science collaborations. Nature 2021;594:316–9. https://doi.org/10.1038/d41586-021-01570-2.

[3] Kadakia KT, Beckman AL, Ross JS, Krumholz HM. Leveraging open science to accelerate research. N Engl J Med 2021;384:e61. https://doi.org/10.1056/NEJMp2034518.

[4] Kodama K, et al. Expression-based genome-wide association study links the receptor CD44 in adipose tissue with type 2 diabetes. Proc Natl Acad Sci 2012;109:7049–54. https://doi.org/10.1073/pnas.1114513109.

[5] Fleischer JG, et al. Predicting age from the transcriptome of human dermal fibroblasts. Genome Biol 2018;19:221. https://doi.org/10.1186/s13059-018-1599-6.

[6] Caliskan A, Crouch SAW, Giddins S, Dandekar T, Dangwal S. Progeria and aging - omics based comparative analysis. Biomedicines 2022;10:2440.

[7] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10. https://doi.org/10.1093/nar/30.1.207.

[8] Barrett T, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 2013;41:D991–5. https://doi.org/10.1093/nar/gks1193.

[9] Baker M. Gene data to hit milestone. Nature 2012;487:282–3. https://doi.org/10.1038/487282a.

[10] Sen A. Metadata management: past, present and future. Decis Support Syst 2004;37:151–73. https://doi.org/10.1016/S0167-9236(02)00208-7.

[11] Greenberg J, Sutton S, Campbell DG. Metadata: a fundamental component of the semantic web. Bull Am Soc Inf Sci Technol 2003;29:16–8. https://doi.org/10.1002/bult.282.

[12] Berners-Lee, T., Hendler, J. & Lassila, O. in Scientific American (⟨https://www.scientificamerican.com/article/the-semantic-web/⟩, 2001).

[13] Hitzler P. A review of the semantic web field. Commun ACM 2021;64:76–83. https://doi.org/10.1145/3397512.

[14] Berners-Lee, T. Semantic Web - XML2000 <⟨https://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html⟩> (w3.org, 2000).

[15] Prud'hommeaux, E. & Seaborne, A. *SPARQL Query Language for RDF*, <⟨https://www.w3.org/TR/rdf-sparql-query/⟩> (2008).

[16] Prud'hommeaux, E. & Seaborne, A. *SPARQL Query Language for RDF - W3C Candidate Recommendation 14 June 2007*, <⟨https://www.w3.org/TR/2007/CR-rdf-sparql-query-20070614/⟩> (2007).

[17] Hogan, A. The Semantic Web: Two Decades On. *Semantic Web Journal* <⟨https://semantic-web-journal.net/content/semantic-web-two-decades-0⟩> (2019).

[18] Tompkins VT, Honick BJ, Polley KL, Qin J. MetaFAIR: a metadata application profile for managing research data. Proc Assoc Inf Sci Technol 2021;58:337–45. https://doi.org/10.1002/pra2.461.

[19] Wilkinson MD, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

[20] Rocca-Serra P, et al. The FAIR cookbook - the essential resource for and by FAIR doers. Sci Data 2023;10:292. https://doi.org/10.1038/s41597-023-02166-3.

[21] Jati PHP, Lin Y, Nodehi S, Cahyono DB, van Reisen M. FAIR versus open data: a comparison of objectives and principles. Data Intell 2022;4:867–81. https://doi.org/10.1162/dint_a_00176.

[22] Berners-Lee, T. *Linked Data*, <⟨https://www.w3.org/DesignIssues/LinkedData.html⟩> (2006).

[23] Berners-Lee, T. *5 Star Linked Data* , <⟨https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data⟩> (2013).

[24] Ulrich H, et al. Understanding the nature of metadata: systematic review. J Med Internet Res 2022;24:e25440. https://doi.org/10.2196/25440.

[25] Furner J. Definitions of "metadata": a brief survey of international standards. J Assoc Inf Sci Technol 2020;71:E33–42. https://doi.org/10.1002/asi.24295.

[26] Habermann T. Metadata life cycles, use cases and hierarchies. Geosciences 2018; 8:179.

[27] Rajesh A, et al. Improving the completeness of public metadata accompanying omics studies. Genome Biol 2021;22:106. https://doi.org/10.1186/s13059-021-02332-z.

[28] Miller GA, Fellbaum C. WordNet then and now. Lang Resour Eval 2007;41:209–14.

[29] Miller GA. WordNet: a lexical database for English. Commun ACM 1995;38:39–41. https://doi.org/10.1145/219717.219748.

[30] Lam, K.N., Al Tarouti, F. & Kalita, J. in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers)*. 106–111 (Association for Computational Linguistics).

[31] Mayer G, et al. Controlled vocabularies and ontologies in proteomics: overview, principles and practice. Biochim Et Biophys Acta (BBA) - Proteins Proteom 2014; 1844:98–107. https://doi.org/10.1016/j.bbapap.2013.02.017.

[32] Ahmad A, Justo JLB, Feng C, Khan AA. The impact of controlled vocabularies on requirements engineering activities: a systematic mapping study. Appl Sci 2020; 10:7749.

[33] Jackson R, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. doi:10.1093/database/baab069 Database 2021;2021: baab069. doi:10.1093/database/baab069.

[34] Whetzel PL, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 2011;39:W541–5. https://doi.org/10.1093/nar/gkr469.

[35] Hartl, N. *Funding of new NFDI Basic Services for Persistent Identifiers and Terminologies*, <⟨https://www.nfdi.de/funding-of-new-nfdi-basic-services-for-persistent-identifiers-and-terminologies/?lang=en⟩> (2023).

[36] Hartl, N. *NFDI publishes statement on the EU Data Act*, <⟨https://www.nfdi.de/nfdi-publishes-statement-on-the-eu-data-act/?lang=en⟩> (2022).

[37] Rocca-Serra, P. et al., <⟨https://doi.org/10.5281/zenodo.6783564⟩> (2022).

[38] Ioannidis JPA. Meta-research: the art of getting it wrong. Res Synth Methods 2010;1:169–84. https://doi.org/10.1002/jrsm.19.

[39] Abraham R, Schneider J, vom Brocke J. Data governance: a conceptual framework, structured review, and research agenda. Int J Inf Manag 2019;49: 424–38. https://doi.org/10.1016/j.ijinfomgt.2019.07.008.

[40] Gribbin J. in 13.8: The Quest to Find the True Age of the Universe and the Theory of Everything Ch. 5. 31.415 Prehistory: Galaxies and the Universe at large. Yale University Press; 2016. p. 115–38.

[41] van Helvoirt, S. & Weigand, H. in *Open and Big Data Management and Innovation.* (eds Marijn Janssen et al.) 160–172 (Springer International Publishing).

[42] Tse, D., Chow, C. k, Ly, T. p, Tong, C.Y. & Tam, K.W. in 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). 1632–1636.

[43] Shabani M. The Data Governance Act and the EU's move towards facilitating data sharing. Mol Syst Biol 2021;17:e10229. https://doi.org/10.15252/msb.202110229.

[44] Zimmerman MD, et al. In: Wayne FAnderson, editor. in Structural Genomics and Drug Discovery: Methods and Protocols. New York: Springer; 2014. p. 1–25.

[45] Buttner, P. et al. in ⟨https://www.ahima.org/media/pmcb0fr5/healthcare-data-governance-practice-brief-final.pdf⟩ (ed AHIMA) (AHIMA, 2022).

[46] Fränti P, Mariescu-Istodor R. Soft precision and recall. Pattern Recognit Lett 2023;167:115–21. https://doi.org/10.1016/j.patrec.2023.02.005.

[47] Ting KM. In: Claude Sammut, Webb) Geoffrey I, editors. in *Encyclopedia of Machine Learning*. Springer US; 2010. 781-781.

[48] Musen MA, et al. Modeling community standards for metadata as templates makes data FAIR. Sci Data 2022;9:696. https://doi.org/10.1038/s41597-022-01815-3.

[49] Barry E, Bannister F. Barriers to open data release: a view from the top. Inf Polity 2014;19:129–52. https://doi.org/10.3233/IP-140327.

[50] Blanco-Melo D, et al. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. e1039 Cell 2020;181:1036–45. https://doi.org/10.1016/j.cell.2020.04.026.

[51] Han Y, et al. Identification of SARS-CoV-2 inhibitors using lung and colonic organoids. Nature 2021;589:270–5. https://doi.org/10.1038/s41586-020-2901-9.

[52] Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013; 29:15–21. https://doi.org/10.1093/bioinformatics/bts635.

[53] Frankish A, et al. GENCODE 2021. Nucleic Acids Res 2021;49:D916–23. https://doi.org/10.1093/nar/gkaa1087.

[54] Frankish A, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47:D766–73. https://doi.org/10.1093/nar/gky955.

[55] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinforma 2011;12:323. https://doi.org/10.1186/1471-2105-12-323.

[56] Soneson C, Love M, Robinson M. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved]. F1000Research 2016;4. https://doi.org/10.12688/f1000research.7563.2.

[57] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550. https://doi.org/10.1186/s13059-014-0550-8.

[58] Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics 2019; 35:2084–92. https://doi.org/10.1093/bioinformatics/bty895.

[59] Kolde R. pheatmap: Pretty Heatmaps (R package, ⟨https://CRAN.R-project.org/package=pheatmap⟩, 2019).

[60] Brockdorff N, et al. Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. Nature 1991;351:329–31. https://doi.org/10.1038/351329a0.

[61] Nafian Dehkordi S, Khani F, Hassani SN, Baharvand H, Soleimanpour-lichaei HR. The contribution of Y chromosome genes to spontaneous differentiation of human embryonic stem cells into embryoid bodies in vitro. Cell J 2021;23:40–50. https://doi.org/10.22074/cellj.2021.7145.

[62] Dasari VK, et al. Expression analysis of y chromosome genes in human prostate cancer. J Urol 2001;165:1335–41. https://doi.org/10.1016/S0022-5347(01)69895-1.

[63] Park Y, et al. Identification of human gene research articles with wrongly identified nucleotide sequences. Life Sci Alliance 2022;5:e202101203. https://doi.org/10.26508/lsa.202101203.

[64] Park Y, et al. Human gene function publications that describe wrongly identified nucleotide sequence reagents are unacceptably frequent within the genetics literature. *bioRxiv*, 2021 2007 2029 453321 2021. https://doi.org/10.1101/2021.07.29.453321.

[65] Toker L, Feng M, Pavlidis P. Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies [version 2; peer review: 2 approved, 1 approved with reservations]. F1000Research 2016;5. https://doi.org/10.12688/f1000research.9471.2.

[66] Qu C, et al. Cost–effective prediction of gender-labeling errors and estimation of gender-labeling error rates in candidate-gene association studies. Front Genet 2011;2. https://doi.org/10.3389/fgene.2011.00031.

[67] Abeysooriya M, Soria M, Kasu MS, Ziemann M. Gene name errors: lessons not learned. PLOS Comput Biol 2021;17:e1008984. https://doi.org/10.1371/journal.pcbi.1008984.

[68] Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature. Genome Biol 2016;17:177. https://doi.org/10.1186/s13059-016-1044-7.

[69] Cioffi A, et al. Identifying and correcting invalid citations due to DOI errors in Crossref data. doi:10.1007/s11192-022-04367-w Scientometrics 2022;127: 3593–612. doi:10.1007/s11192-022-04367-w.

[70] Brembs B. Prestigious science journals struggle to reach even average reliability. Front Hum Neurosci 2018;12. https://doi.org/10.3389/fnhum.2018.00037.

[71] Krachunov M, Nisheva M, Vassilev D. Machine learning models for error detection in metagenomics and polyploid sequencing. Data *Inf* 2019;10.

[72] Kunis S, et al. MDEmic: a metadata annotation tool to facilitate management of FAIR image data in the bioimaging community. Nat Methods 2021;18:1416–7. https://doi.org/10.1038/s41592-021-01288-z.

[73] Baker M. 1,500 scientists lift the lid on reproducibility. Nature 2016;533:452–4. https://doi.org/10.1038/533452a.

[74] Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2: e124. https://doi.org/10.1371/journal.pmed.0020124.

[75] Ioannidis JPA, Munafò MR, Fusar-Poli P, Nosek BA, David SP. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends Cogn Sci 2014;18:235–41. https://doi.org/10.1016/j.tics.2014.02.010.

76 Mashoufi M, Ayatollahi H, Khorasani-Zavareh AD. Data quality assessment in emergency medical services: what are the stakeholders' perspectives? Perspect Health Inf Manag 2019;16:1c.

[77] Perumal N, et al. Anthropometric data quality assessment in multisurvey studies of child growth. Am J Clin Nutr 2020;112:806S–15S. https://doi.org/10.1093/ajcn/nqaa162.

[78] Liaw S-T, et al. Quality assessment of real-world data repositories across the data life cycle: a literature review. J Am Med Inform Assoc 2021;28:1591–9. https://doi.org/10.1093/jamia/ocaa340.

[79] Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLOS Med 2009;6:e1000097. https://doi.org/10.1371/journal.pmed.1000097.

[80] Tute E, Scheffner I, Marschollek M. A method for interoperable knowledge-based data quality assessment. BMC Med Inform Decis Mak 2021;21:93. https://doi.org/10.1186/s12911-021-01458-1.

[81] Yao B, Kang H, Gong Y. Data quality assessment of narrative medication error reports. Stud Health Technol Inf 2019;265:101–6. https://doi.org/10.3233/shti190146.

[82] Caliskan D, et al. First steps to evaluate an NLP tool's medication extraction accuracy from discharge letters. Stud Health Technol Inf 2021;278:224–30. https://doi.org/10.3233/shti210073.

[83] Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv: 2307 09009v2 2023. https://doi.org/10.48550/arXiv.2307.09009.

[84] Hicks SA, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep 2022;12:5979. https://doi.org/10.1038/s41598-022-09954-8.

[85] Schmedes SE, King JL, Budowle B. Correcting inconsistencies and errors in bacterial genome metadata using an automated curation tool in excel (AutoCurE). Front Bioeng Biotechnol 2015;3. https://doi.org/10.3389/fbioe.2015.00138.

[86] Crandall ED, et al. Importance of timely metadata curation to the global surveillance of genetic diversity. Conserv Biol 2023;n/a:e14061. https://doi.org/10.1111/cobi.14061.

[87] Goudey B, Geard N, Verspoor K, Zobel J. Propagation, detection and correction of errors using the sequence database network. Brief Bioinforma 2022;23:bbac416. https://doi.org/10.1093/bib/bbac416.

[88] Ridzuan F, Wan Zainon WMN. A review on data cleansing methods for big data. Procedia Comput Sci 2019;161:731–8. https://doi.org/10.1016/j.procs.2019.11.177.

[89] Chai CP. The importance of data cleaning: three visualization examples. Chance 2020;33:4–9. https://doi.org/10.1080/09332480.2020.1726112.

[90] Panse C, Trachsel C, Türker C. Bridging data management platforms and visualization tools to enable ad-hoc and smart analytics in life sciences. J Integr Bioinforma 2022;19. https://doi.org/10.1515/jib-2022-0031.

[91] Beretta V, et al. A user-centric metadata model to foster sharing and reuse of multidisciplinary datasets in environmental and life sciences. Comput Geosci 2021;154:104807. https://doi.org/10.1016/j.cageo.2021.104807.

[92] Kamdar MR, Musen MA. An empirical meta-analysis of the life sciences linked open data on the web. Sci Data 2021;8:24. https://doi.org/10.1038/s41597-021-00797-y.

[93] McCrae, J.P. et al. *The Linked Open Data Cloud*, <⟨http://cas.lod-cloud.net/⟩> (2023).

[94] Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet 2020;21:615–29. https://doi.org/10.1038/s41576-020-0257-5.

[95] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015;33:495–502. https://doi.org/10.1038/nbt.3192.

[96] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20. https://doi.org/10.1038/nbt.4096.

[97] Stuart T, et al. Comprehensive Integration of Single-Cell Data. e1821 Cell 2019; 177:1888–902. https://doi.org/10.1016/j.cell.2019.05.031.

[98] Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol 2019; 20:296. https://doi.org/10.1186/s13059-019-1874-1.

[99] Hao Y, et al. Integrated analysis of multimodal single-cell data. e3529 Cell 2021; 184:3573–87. https://doi.org/10.1016/j.cell.2021.04.048.

[100] Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods 2016;13:966–7. https://doi.org/10.1038/nmeth.4077.

[101] Türei D, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol 2021;17:e9923. https://doi.org/10.15252/msb.20209923.

[102] Kozareva V, et al. A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types. Nature 2021;598:214–9. https://doi.org/10.1038/s41586-021-03220-z.

[103] Seering J. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. Article 107, doi: 10.1145/3415178 Proc ACM Hum -Comput Interact 2020;4. Article 107, doi: 10.1145/3415178.

[104] Blanco-Melo D, et al. SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. *bioRxiv*, 2020 2003 2024 004655 2020. https://doi.org/10.1101/2020.03.24.004655.

[105] Ishwarappa, Anuradha J. A brief introduction on big data 5Vs characteristics and hadoop technology. Procedia Comput Sci 2015;48:319–24. https://doi.org/10.1016/j.procs.2015.04.188.

[106] García Lozano M, et al. Veracity assessment of online data. Decis Support Syst 2020;129:113132. https://doi.org/10.1016/j.dss.2019.113132.

[107] de Souza Granha RGD. In: Sherif Sakr, Albert YZomaya, editors. *Encyclopedia of Big Data Technologies*. Springer International Publishing; 2019. p. 913–7.

[108] Azeroual O, Fabre R. Processing big data with apache hadoop in the current challenging era of COVID-19. Big Data Cogn Comput 2021;5:12.

[109] Andrešić, D., Šaloun, P. & Anagnostopoulos, I. in *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*. 1–5. (2023).

[110] Kumar, D. & Li, S. in 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA). 1–2.

[111] Kalla, D., Samaah, F., Kuraku, S. & Smith, N. Phishing Detection Implementation Using Databricks and Artificial Intelligence. SSRN Electronic Journal **185**, doi: 10.2139/ssrn.4452780 (2023).

[112] Van Noorden R. Science publishing: the trouble with retractions. Nature 2011; 478:26–8. https://doi.org/10.1038/478026a.

[113] Barbour V, Bloom T, Lin J, Moylan E. Amending published articles: time to rethink retractions and corrections? [version 1; peer review: 2 approved with reservations]. F1000Research 2017;6. https://doi.org/10.12688/f1000research.13060.1.

[114] Allison DB, Brown AW, George BJ, Kaiser KA. Reproducibility: a tragedy of errors. Nature 2016;530:27–9. https://doi.org/10.1038/530027a.

[115] Klingner CM, et al. Research data management and data sharing for reproducible research—results of a community survey of the german national research data infrastructure initiative. neuroscience eneuro 2023;10(2):ENEURO.0215-0222. https://doi.org/10.1523/ENEURO.0215-22.2023.

[116] Harrow J, Hancock J, Community E-E, Blomberg N. ELIXIR-EXCELERATE: establishing Europe's data infrastructure for the life science research of the future. EMBO J 2021;40:e107409. https://doi.org/10.15252/embj.2020107409.

[117] Sayers EW, O'Sullivan C, Karsch-Mizrachi I. In: Using GenBank and SRA David Edwards,editor. Plant Bioinformatics: Methods and Protocols. US: Springer; 2022. p. 1–25.