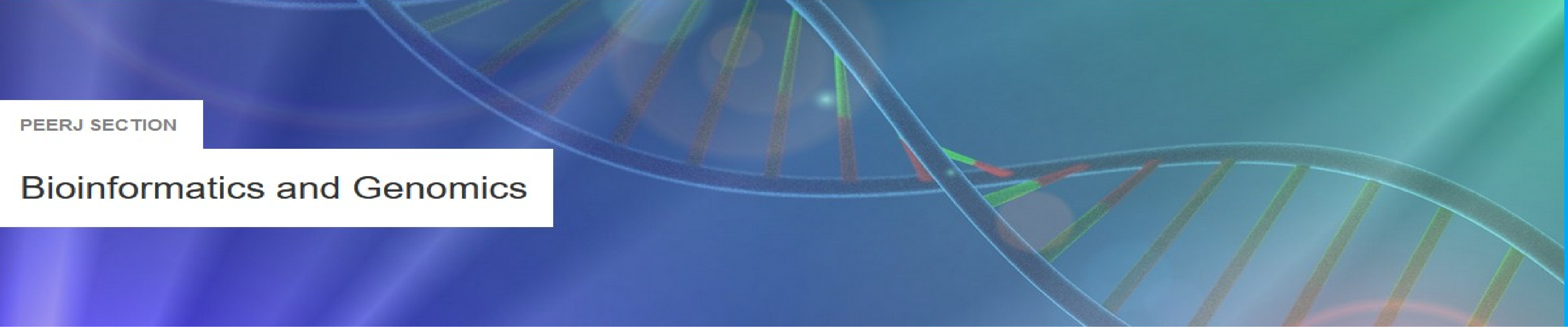


PeerJ
Life & Environment

The reuse of public datasets in the life sciences: potential risks and rewards

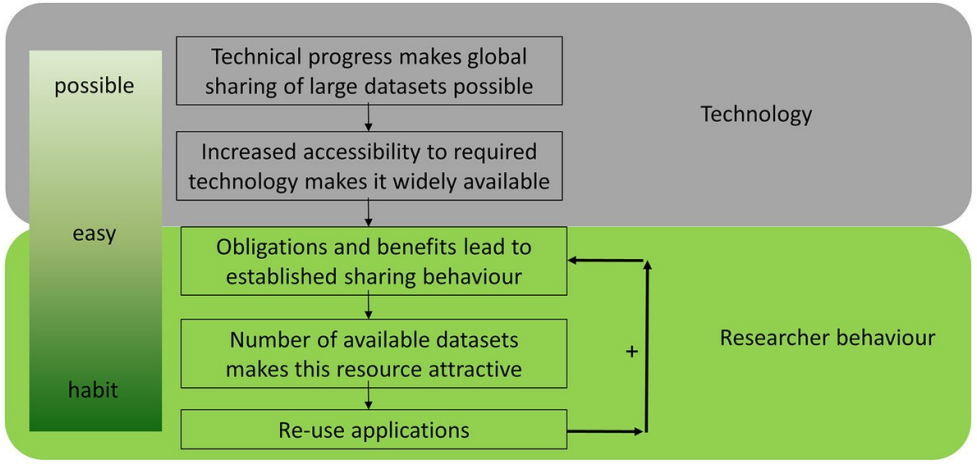
Katharina Sielemann, Alenka Hafner and Boas Pucker

Seminário de Afinidades em Genômica e Bioinformática - Terceiro encontro de 2024

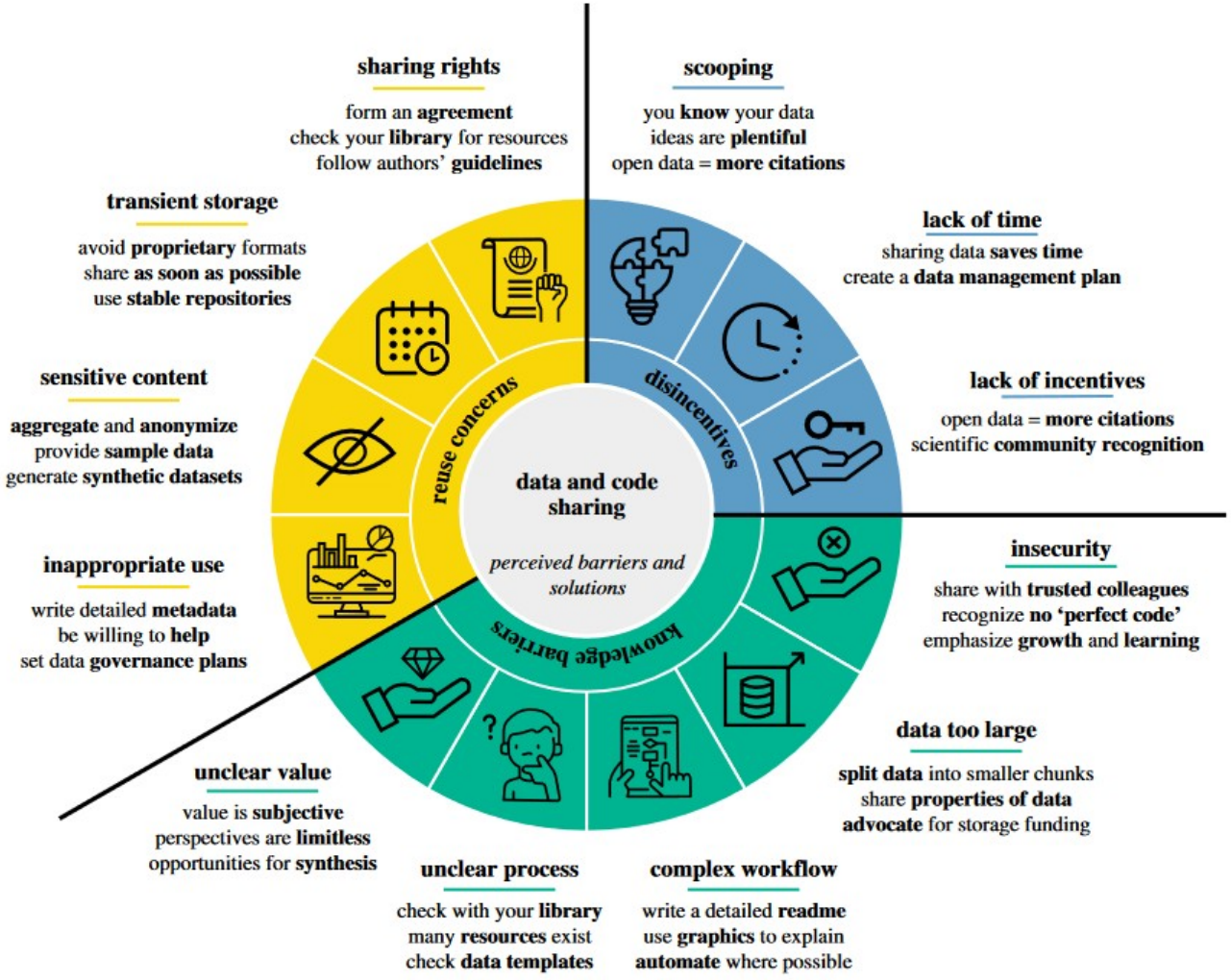
28/08/2024

Big data

- A expansão do “big data” resultou em novas análises para diversos tipos de dados já publicados
- Desenvolvimento de ferramentas de bioinformática para extração de dados já disponíveis:
  - Testes de ferramentas de bioinformática na identificação de padrões
  - Fontes de conjuntos de dados para treinamento de algoritmos de aprendizado de máquinas



Fonte: Sielemann et al. (2020)



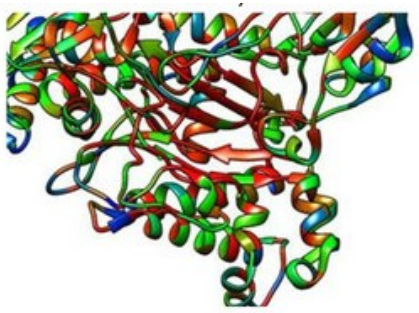
Fonte: Gomes et al. (2022)

Tipos de dados

- Dados primários são diretos e obtidos experimentalmente (sequências, metodologias, grafos, anotações, códigos, medidas, padrões, estruturas, bases de dados primárias)

```
>seq1
ATCGTTTAGCTAGACCTGATG
ATCCGATCGATTACGTG
>seq2
GACACGATCGTCAGAAATGCA
GTC
>seq3
ACGACAAATCATCTCC
```

```
>seq1
MYVRANQEFFK
>seq2
WTSMADCHLKV
>seq3
MGPKLHIGRQEFKLIHYWN
NNG
```



```
for ID in transcripts_per_genes[ gene ]:
        counter = 0
        for element in transcripts[ ID ]:
                if element['type'] == "CDS":
                        counter += element['end']-
element['start']
        CDS_len_per_transcript.append( { 'id': ID, 'len': counter } )
repr_trans = sorted( CDS_len_per_transcript, key=itemgetter('len') )[-1]
repr_transcripts.update( { repr_trans['id']: transcripts[ repr_trans['id'] ] } )
```

PDBe
Protein Data Bank in Europe

- Dados derivados são dados de meta-análises e processados (publicações, bases de dados curadas)

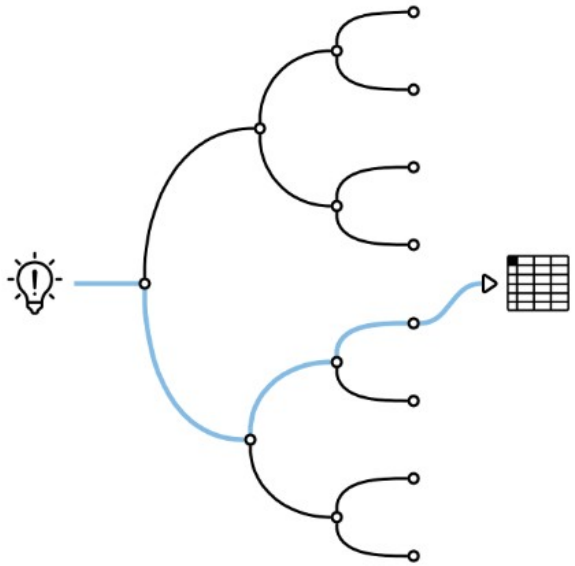
InterPro

KEGG Kyoto Encyclopedia of Genes and Genomes

Fonte: Sielemann et al. (2020)

Potenciais

- O estabelecimento de normas de compartilhamento de dados devem ser introduzidas por obrigações nas publicações, incluindo descrição e padronização



Fonte: Borghi e Van Gulick. (2022)

	Project Planning	**Data Collection and Analysis**	**Data Publication and Sharing**
Data Management Activities	• Data Management Planning	• Saving and backing up files • Organizing files • Formatting and describing data according to standards • Maintaining documentation and metadata	• Preserving data and other materials (e.g. reagents, code) • Assigning persistent identifiers
Open Science Activities	• Planning for open (e.g. including data sharing in consent forms)	• Using open source tools • Using transparent methods and protocols	• Sharing data • Publishing research reports openly (e.g. open access publishing).
Reproducibility-Related Activities	• Preregistering study aims and methods • Using appropriate research designs (including sufficient statistical power)	• Preventing methodological issues (e.g. p-hacking, HARKing) • Implementing quality control measures	• Preventing publication bias • Following reporting guidelines

Fonte: Borghi e Van Gulick. (2022)

Potenciais

- A padronização exige precauções:
  - A integridade da fonte deve ser verificada
  - Possíveis vieses devem ser analisados
  - Todas as informações relevantes devem estar disponíveis (metadados)
  - A integração de dados só pode ser realizada com dados semelhantes
  - A qualidade deve ser alta (Phred scores, comprimento e pareamento de sequências, contig/scaffold N50, etc)

Name	Explanation	Prevention
Allocation bias	Systematic difference in the assignment of participants to the treatment and control group in a clinical trial. For example, the investigator knows or can predict which intervention the next eligible patient is supposed to receive due to poorly concealed randomization.	- Randomization with allocation concealment
Attrition bias	Attrition occurs when participants leave during a study that aims to explore the effect of continuous exposure (dropouts or withdrawal). For example, more dropouts of patients randomized to an aggressive cancer treatment.	- Good investigator–patient communication - Accessibility of clinics - Incentives to continue
Confounding bias	An artificial association between an exposure and an outcome because another variable is related to both the exposure and outcome. For example, lung cancer risk in coffee drinkers is evaluated, ignoring smoking status (smoking is associated with both coffee drinking and cancer). A challenge is that many confounders are unknown and/or not measured.	- Randomization (can address unmeasured confounders) When randomization is not possible: - Restriction to one level of the confounder - Matching on the levels of the confounder - Stratification and analysis within strata - Propensity score matching
Immortal time bias	Survival beyond a certain time point is necessary in order to be exposed (participants are “immortal” in that time period). For example, discharged patients are analyzed but were included in the treatment group only if they filled a prescription for a drug 90 days after discharge from hospital.	- Group assignment at time zero - Time-dependent analysis may be used
Information bias	Bias that arises from systematic differences in the collection, recall, recording, or handling of information. For example, blood pressure in the treatment arm is measured in the morning and for the control arm in the evening.	- Standardized data collection - Data collection independent from exposure or outcome (e.g., by blinding of intervention status/exposure) - Use of objective measurements
Publication bias	Occurs when only studies with a positive or negative result are published. Affects meta-analyses from systematic reviews and harms evidence-based medicine	- Writing a study protocol and preregistration - Publishing study protocol or registered report - Following reporting guidelines

Fonte: Schwab et al. (2022)

Planning
1. Specify your research question
2. Write and register a study protocol
3. Justify your sample size
4. Write a data management plan
5. Reduce bias

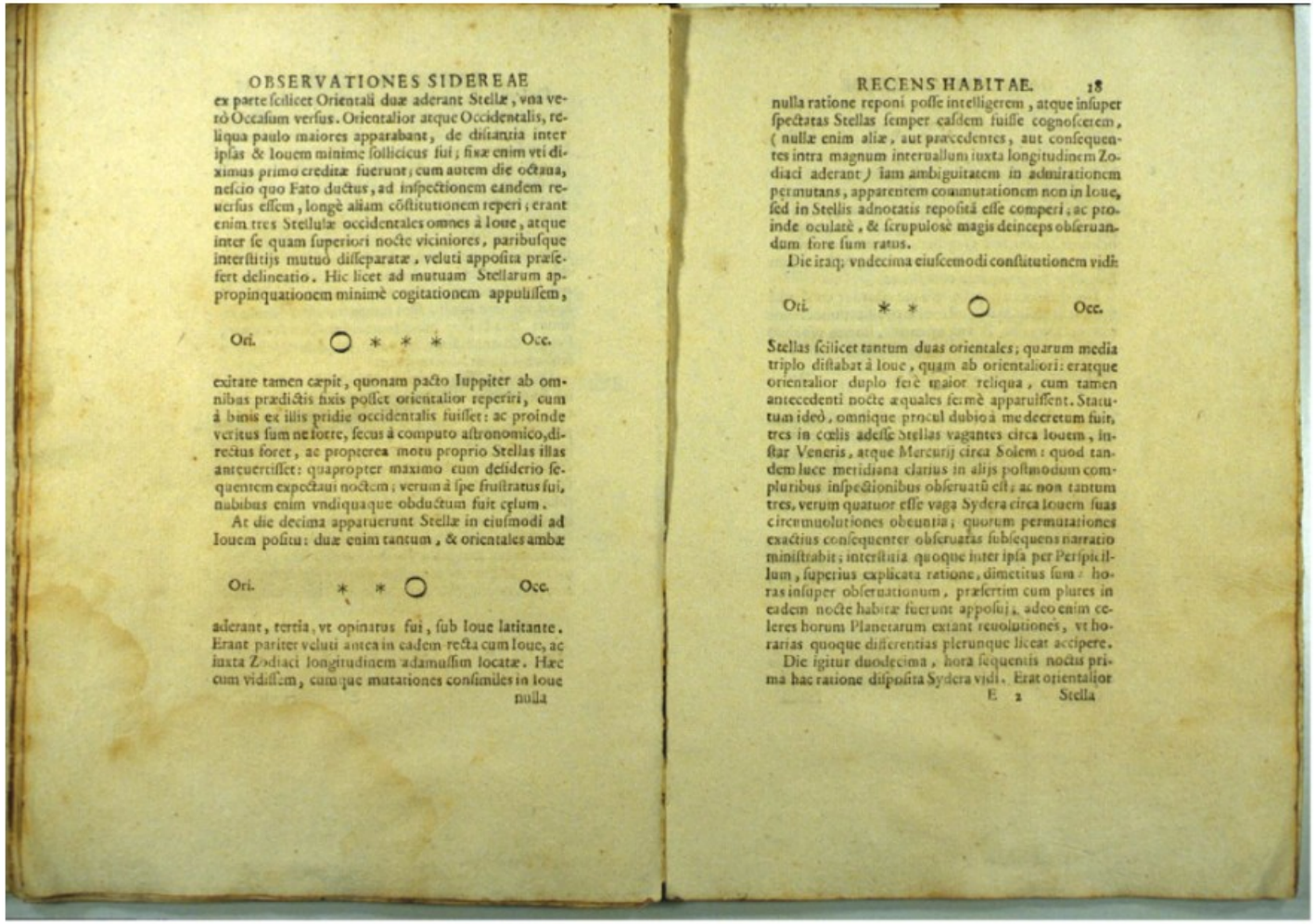
Execution
6. Avoid questionable reseach practices
7. Be cautious with interpretations of statistical significance
8. Make your research open

Reporting
9. Report all findings
10. Follow reporting guidelines

Fonte: Schwab et al. (2022)

Potenciais

- A manutenção de dados favorece também a integração de dados de diferentes fontes:
  - Fornecem análises necessárias para hipóteses que não podem ser testadas por conjuntos únicos de dados
  - A reutilização de dados supera limitações financeiras de muitos grupos menores para o delineamento de experimentos complexos, uma vez que a geração de novos dados pode ser custosa e laborosa



OBSERVATIONES SIDEREAE

ex parte scilicet Orientali duæ aderant Stellæ, vna verò Occasum versus. Orientalior atque Occidentalis, reliqua paulo maiores apparabant, de distantia inter ipsas & Iouem minime sollicitus fui; fixæ enim vti diximus primo creditæ fuerunt; cum autem die octaua, nescio quo Fato ductus, ad inspectionem eandem reuersus essem, longè aliam cõstitutionem reperi; erant enim tres Stellulæ occidentales omnes à Ioue, atque inter se quam superiori nocte viciniores, paribusque interstitijs mutuò disseparatæ, veluti appositæ præfert delineatio. Hic licet ad mutuam Stellarum appropinquationem minimè cogitationem appulissem,

Ori. ○ * * * Occ.

exitare tamen cæpit, quonam pacto Iuppiter ab omnibus prædictis fixis posset orientalior reperiri, cum à binis ex illis pridie occidentalis fuisset: ac proinde veritus sum ne forte, secus à computo astronomico, directus foret, ac propterea motu proprio Stellas illas antecurrisset: quapropter maximo cum desiderio sequentem expectaui noctem; verum à spe frustratus fui, nubibus enim vndiquaque obductum fuit cælum.

At die decima apparuerunt Stellæ in eiusmodi ad Iouem positu: duæ enim tantum, & orientales ambæ

Ori. * * ○ Occ.

aderant, tertia, vt opinatus fui, sub Ioue latitante. Erant pariter veluti antea in eadem recta cum Ioue, ac iuxta Zodiaci longitudinem adamussim locatæ. Hæc cum vidissem, cumque mutationes consimiles in Ioue nulla

RECENS HABITAE. 18

nulla ratione reponi posse intelligerem, atque insuper spectatas Stellas semper easdem fuisse cognoscerem, (nullæ enim aliæ, aut præcedentes, aut consequentes intra magnum interuallum iuxta longitudinem Zodiaci aderant) iam ambiguitatem in admirationem permutans, apparentem commutationem non in Ioue, sed in Stellis adnotatis repositâ esse comperi, ac proinde oculatè, & scrupulosè magis deinceps obseruandum fore sum ratus.

Die itaq; vndecima eiusce modi constitutionem vidi:

Ori. * * ○ Occ.

Stellas scilicet tantum duas orientales; quarum media triplo distabat à Ioue, quam ab orientaliori: eratque orientalior duplo fere maior reliqua, cum tamen antecedenti nocte æquales fermè apparuissent. Statuam ideò, omnique procul dubio à me decretum fuit, tres in cœlis adesse Stellas vagantes circa Iouem, instar Veneris, atque Mercurij circa Solem: quod tandem luce meridiana clarius in alijs postmodum compluribus inspectionibus obseruatũ est: ac non tantum tres, verum quatuor esse vaga Sydera circa Iouem suas circumuolutiones obeuntia; quorum permutationes exactius consequenter obseruatas subsequens narratio ministrabit; interstitia quoque inter ipsa per Perspicillum, superius explicata ratione, dimetitus sum: horas insuper obseruationum, præsertim cum plures in eadem nocte habitæ fuerunt appositi; adeo enim celeres horum Planetarum extant reuolutiones, vt horarias quoque differentias plerunque liceat accipere.

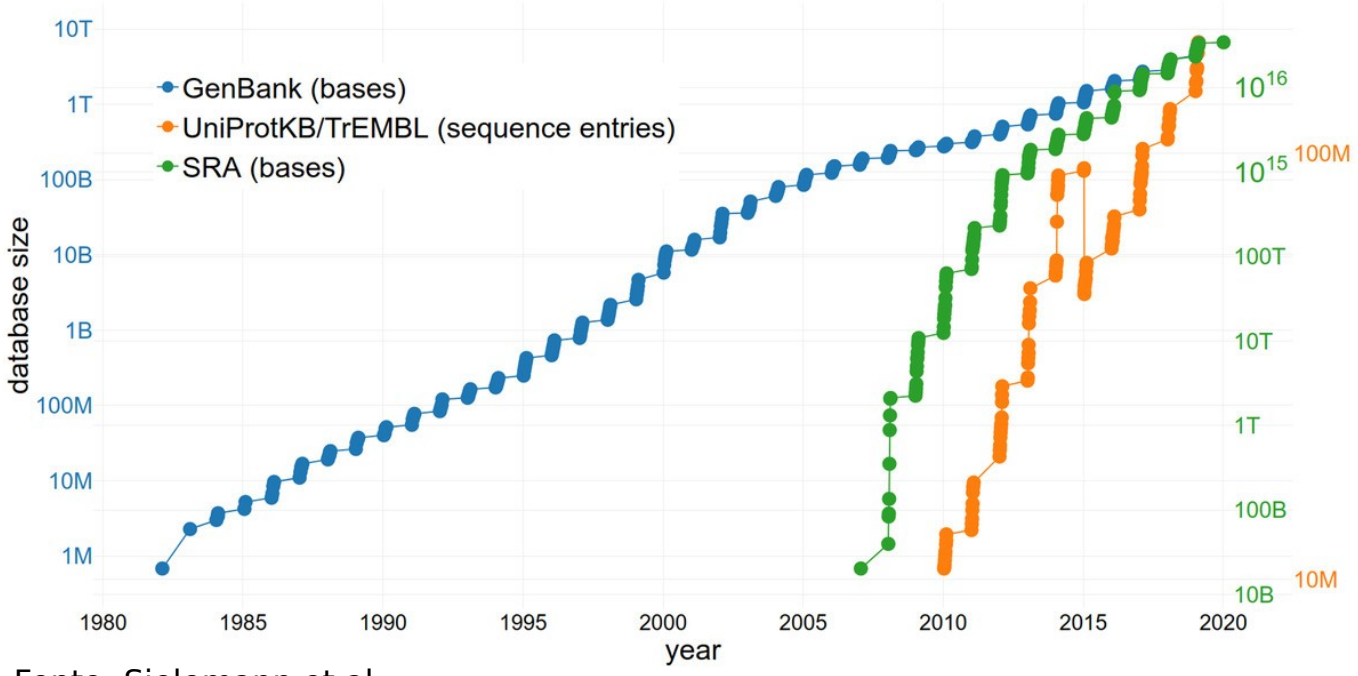
Die igitur duodecima, hora sequentis noctis prima hac ratione disposita Sydera vidi. Erat orientalior

E 2 Stella

Fonte: Goodman et al. (2014)

Potenciais

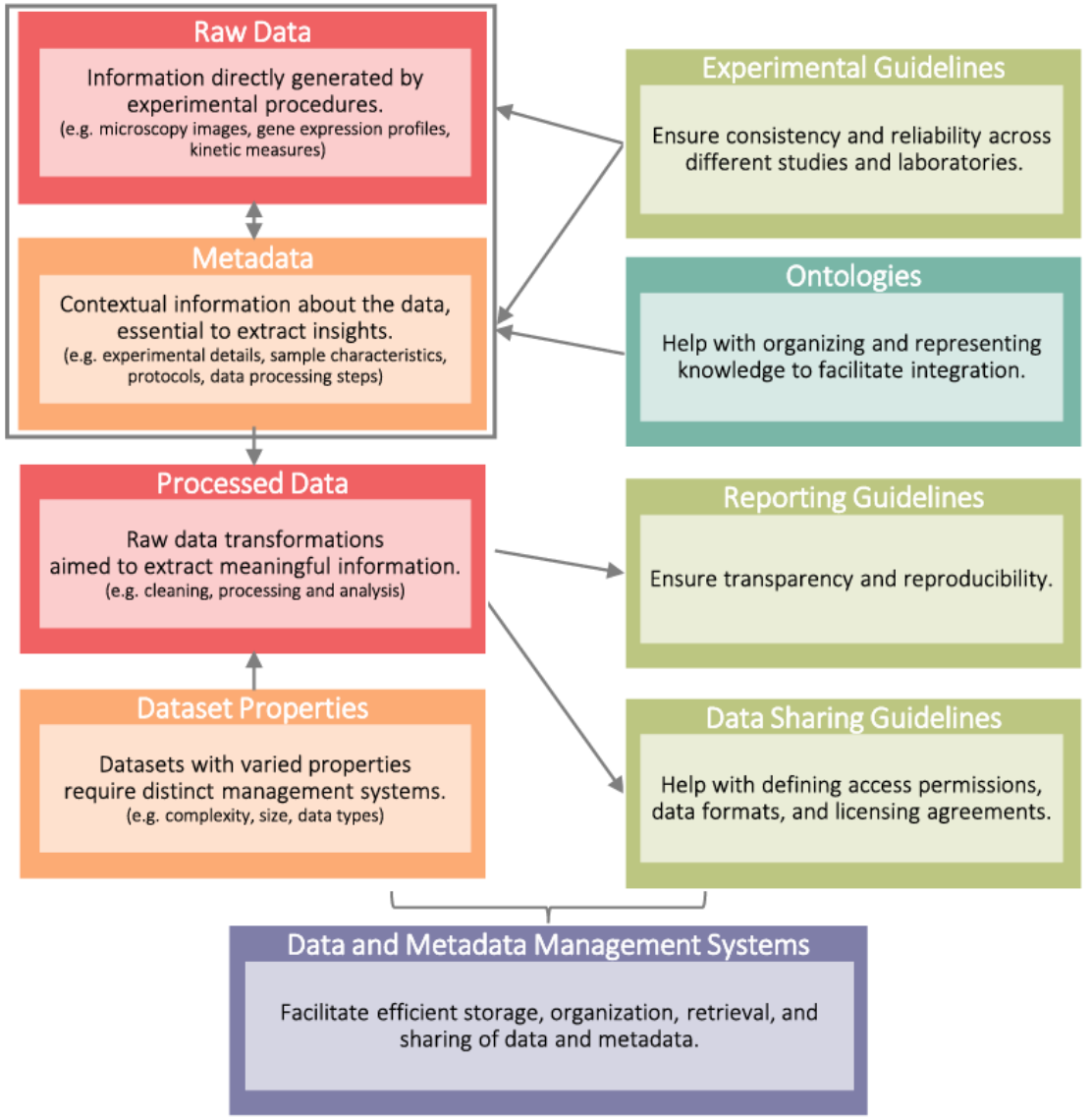
- A disponibilização pública dos dados previne a perda de informação e a geração de conjuntos redundantes e/ou duplicados:
  - Manutenção concisa das bases de dados
  - 20% das publicações de metagenomas entre 2016 e 2019 não são acessíveis ao público



Fonte: Sielemann et al. (2020)

Potenciais

- A reprodutibilidade da reutilização de dados pode ser melhorada pela padronização, documentação e organização de fluxos de trabalho:
  - A integração de diferentes fontes envolve a comparação entre fluxos de trabalho distintos, variáveis indevidamente registradas, apresentações não unificadas, e falta de disponibilidade de dados brutos
  - Exige uma curagem manual
  - Padronização do formato de linguagem e arquivos (XML, FASTA, FASTQ, SAM/BAM)



Fonte: Cunha-Oliveira et al. (2024)

Potenciais

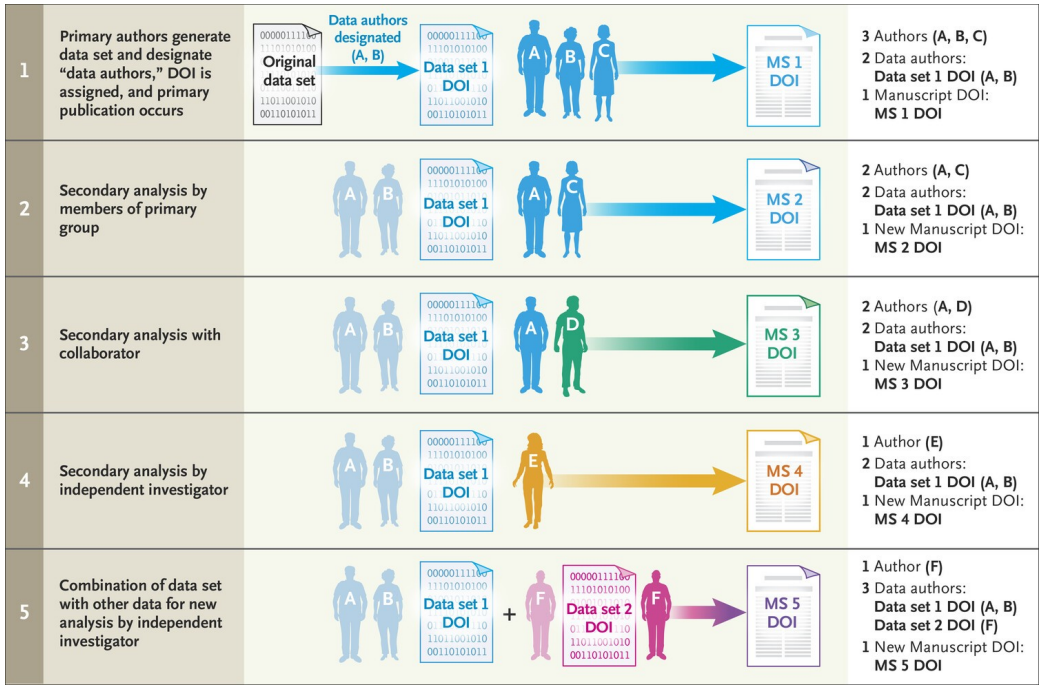
- Existem guias para a padronização do compartilhamento de metadados:
  - Findable Accessible Interoperable Reusable (FAIR)
  - The Transparency and Openness Promotion (TOP)
  - Open Data in a Big Data World (Open Data in a Big Data World, 2016)
  - Beijing Declaration

FAIR Principle	For Infrastructure	For Researchers
Findable	Data and metadata should be easy to find by both humans and computers. In practice, this means that data should be assigned unique and persistent identifiers, described using rich metadata, and registered or indexed in a searchable resource.	Research teams should implement standardized practices related to organizing files (e.g. standardized file naming conventions) so data can be found when needed. When data is made available to others, it should- whenever possible- be uploaded to a repository that assigns a persistent identifier (e.g. DOIs, RRIDs, etc) and describes datasets with standardized metadata. Complete and high quality metadata should be added so data can be discovered and linked to related resources (e.g. related paper DOIs, author ORCIDs).
Accessible	There is a clearly defined method for accessing the data. Data should be retrievable by its identifier using a standardized protocol that is open, free, and universally implementable. Metadata should be accessible even when data is no longer available.	Data is available through a clearly defined process. Members of the research team should be able to access raw data, intermediate products, and other research materials. When data and other materials are made available to others, there should be a clear path to gaining access. The terms by which the data will be made available (e.g. to whom, when, and for what purpose) should be articulated and abided by.
Interoperable	Data should be usable across a range of applications and workflows. Data should use formal, accessible, shared, and broadly applicable models for knowledge representation.	Data should be structured in a standard way so it can be easily combined with other similarly structured datasets. In practice, this means implementing a range of practices such as describing and organizing data (e.g. applying appropriate metadata, maintaining data dictionaries) and saving files in open or non-proprietary file formats.
Reusable	Metadata and data should be described following relevant community standards and have clearly defined conditions for reuse including a machine-readable license	Data should be saved, organized, and described with its future (re)use in mind. A future user may be a member of the research team who is returning to the data after a period of several months or years or another researcher who is (re)using the data for another purpose.

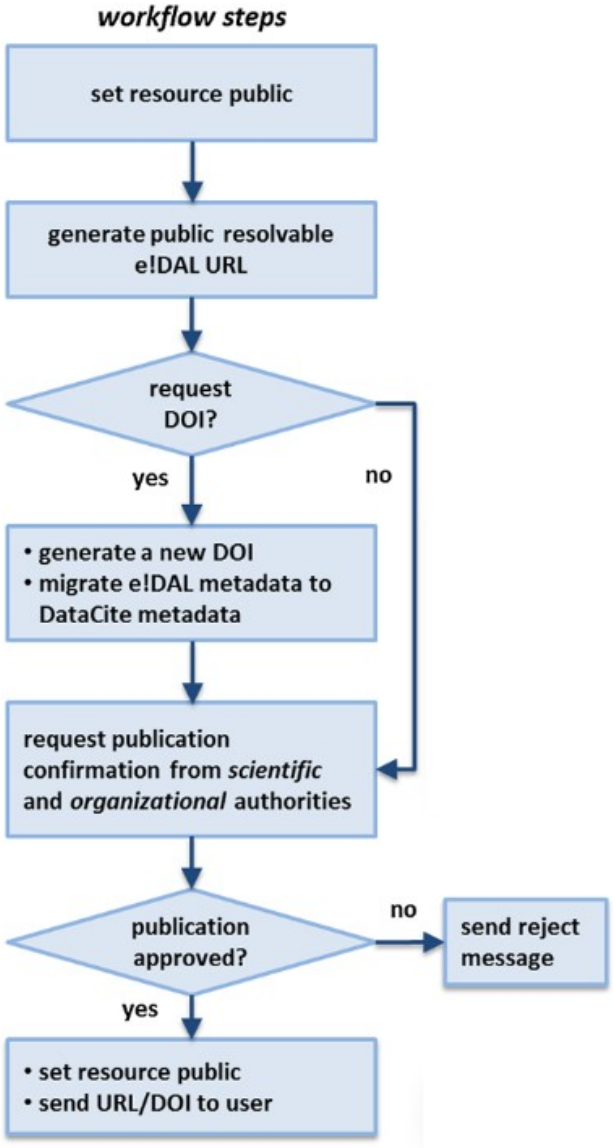
Fonte: Borghi e Van Gulick et al. (2022)

Potenciais

- Autores podem ser beneficiados pelo fornecimento de dados de alta qualidade e com boa anotação:
  - A prática deve fomentar as citações e reconhecimentos adequados
  - Podem resultar em citações adicionais para um mesmo conjunto de dados



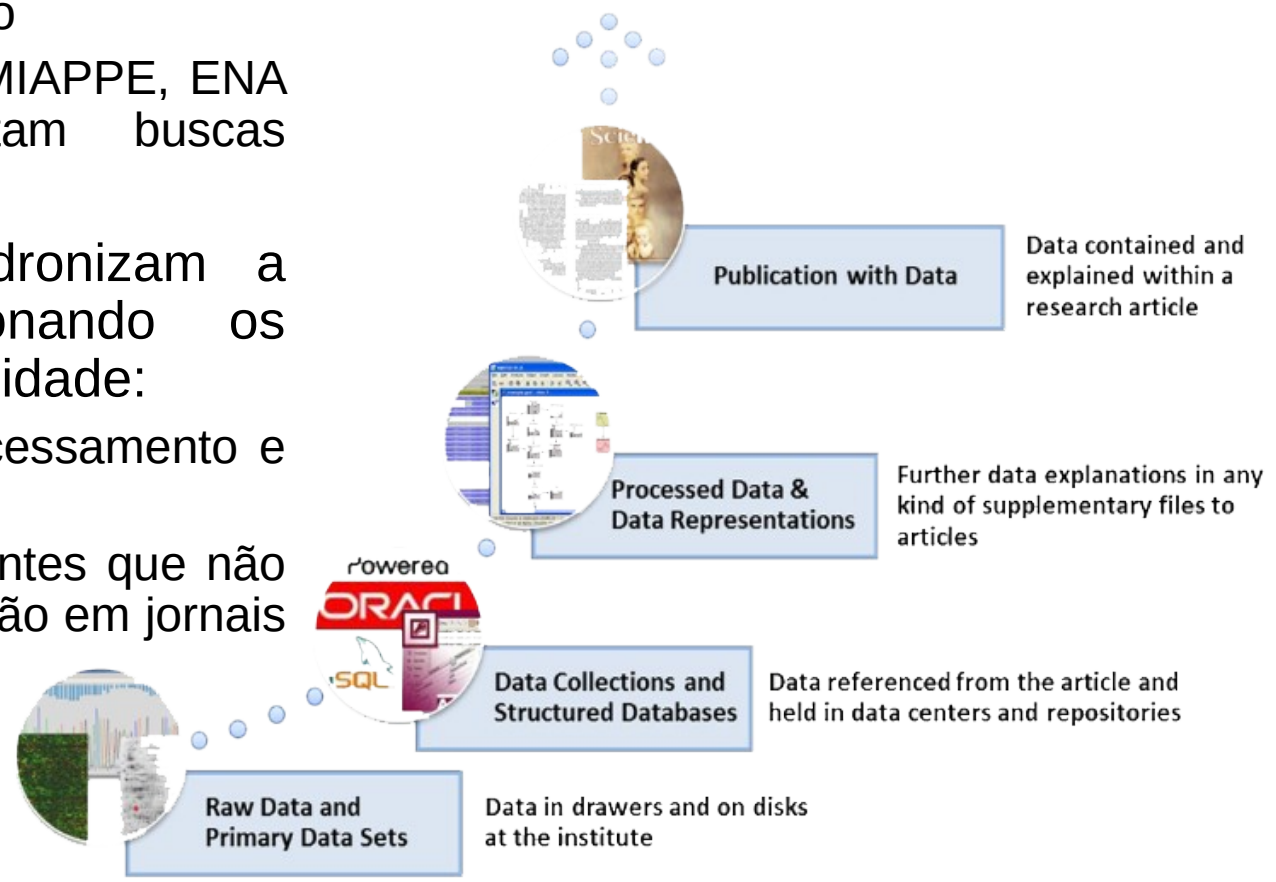
Fonte: Bierer et al. (2017)



Fonte: Arend et al. (2014)

Potenciais

- A padronização exige uma curagem manual:
  - Cada projeto apresenta vocabulário próprio
  - Padronização de metadados (MassIVE, MIAPPE, ENA e SRA) e de vocabulário facilitam buscas automatizadas
- Jornais de dados (*data papers*) padronizam a submissão dos metadados, solucionando os problemas associados ao controle de qualidade:
  - Descrição dos métodos para coleta, processamento e verificação dos dados
  - Permitem a publicação de dados importantes que não são suficientes para compor uma publicação em jornais comuns.



Fonte: Arend et al. (2014)

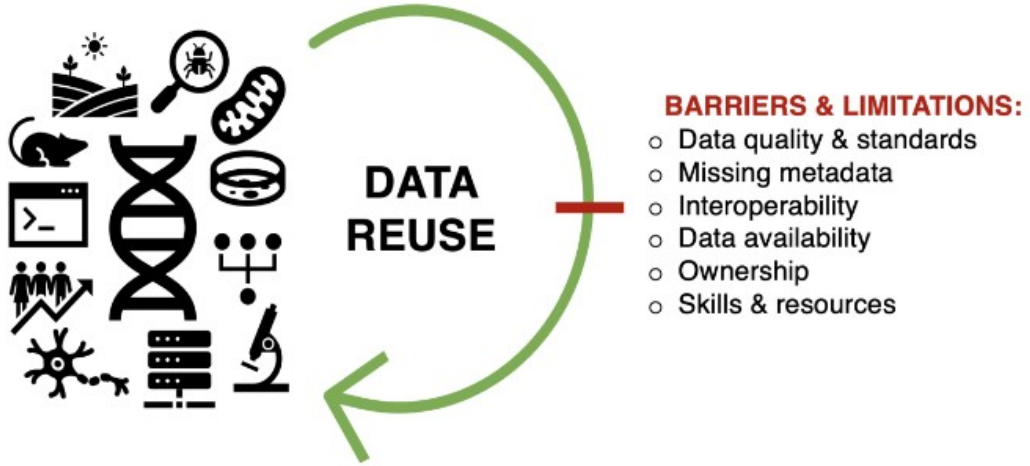
Ética

- Direitos de uso:
  - Reuso justo (para novos propósitos)
  - Reprodução de estudos prévios com dados publicados
  - Reuso injusto (plágio)
- Parasitismo de pesquisa:
  - A reutilização deve gerar novos resultados, e não apenas impulsionar publicações anteriores.
  - *The Research Parasite Awards* premia autores que adequadamente reutilizaram dados

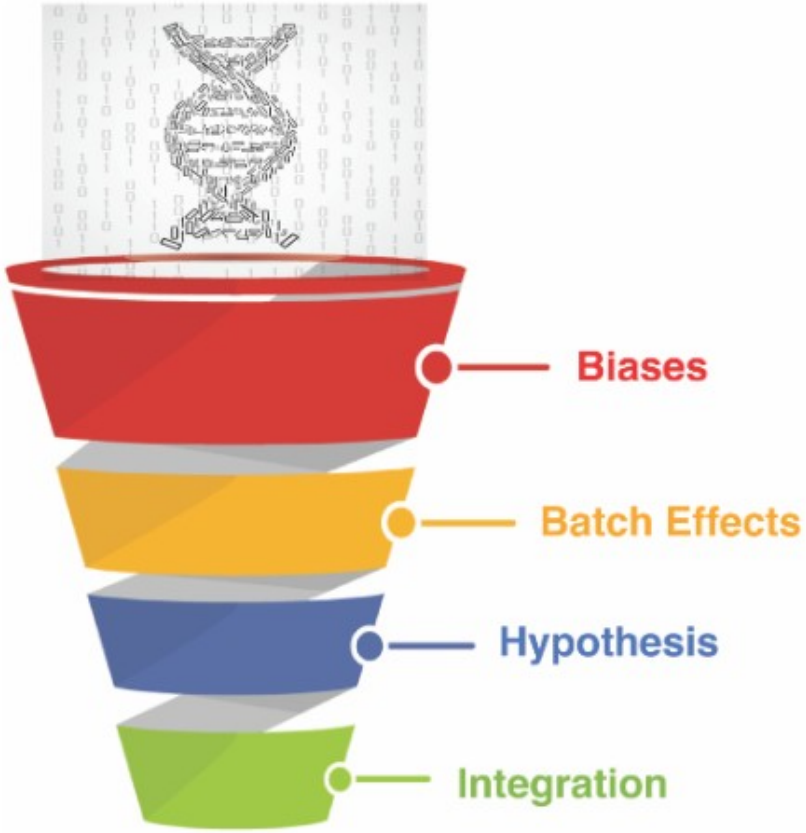
Fonte: Greene Laboratory (2020)

Limitações

- A revisão por pares geralmente não envolve a avaliação dos conjuntos de dados e sua descrição:
  - A qualidade dos dados é questionável quando são de responsabilidade pública
  - Metodologia e design experimental incompletas
  - Propagação de ruídos nos dados com a reutilização
  - Os conjuntos de dados não são normalizados, impedindo a integração



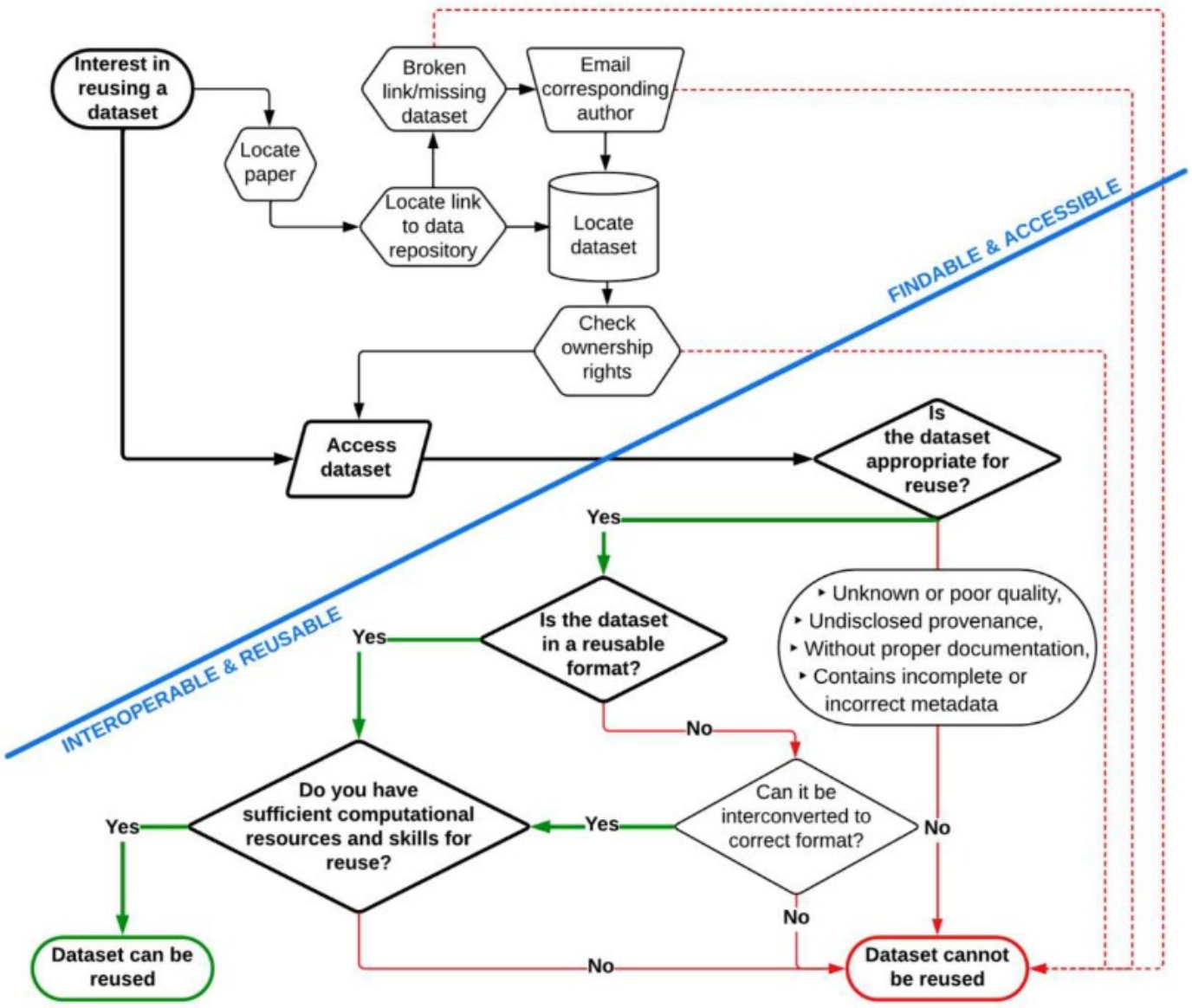
Fonte: Hafner et al. (2024)



Fonte: Raman et al. (2021)

Limitações

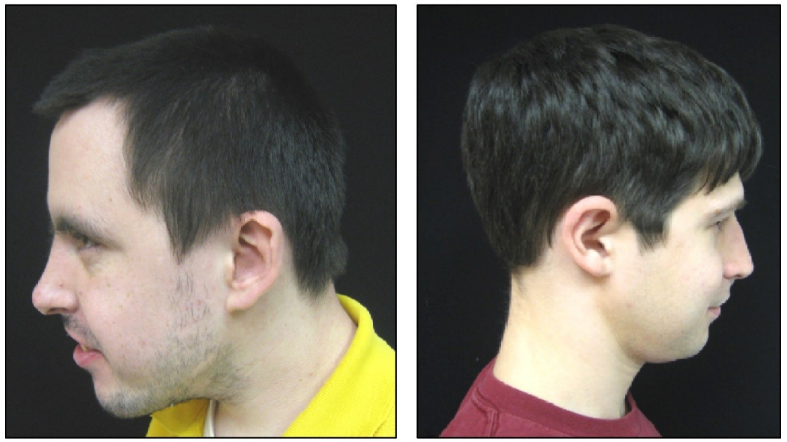
- Sequências podem ser depositadas com identificadores diferentes, impedindo sua integração e escondendo possíveis informações.
- A anotação não é padronizada, possibilitando a inserção de um mesmo conjunto de dados com diferentes identificações
- A montagem de genomas baseada em dados públicos podem ser prejudicadas pela presença de contaminações não identificadas
- Falsos positivos em experimentos podem não ser identificados em dados públicos



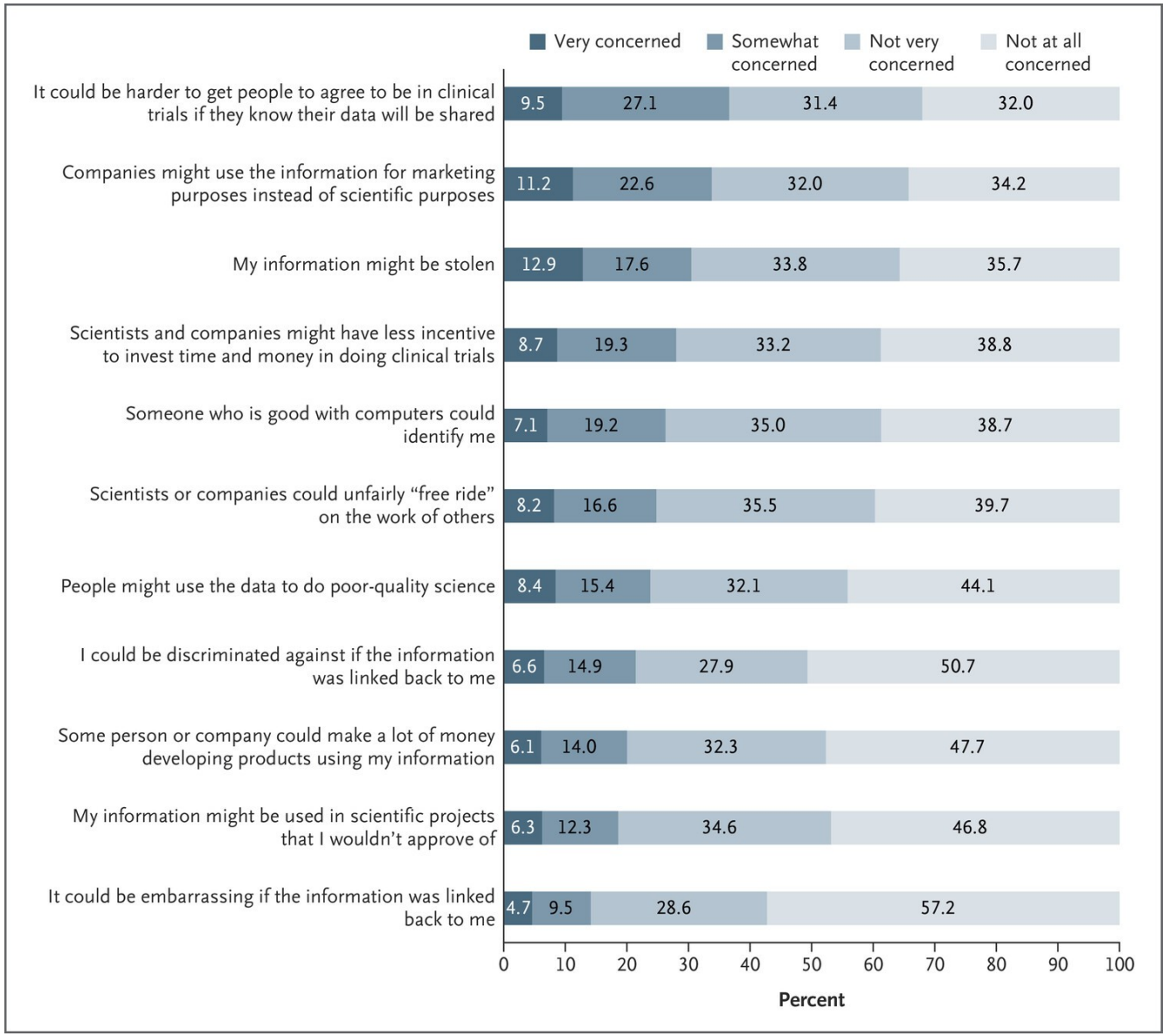
Fonte: Hafner et al. (2024)

Limitações

- Dados clínicos impactam os pacientes envolvidos e devem ser devidamente assegurados para reutilização
  - Pacientes não devem ser identificados
  - Dados genômicos (risco de privacidade para a progênie)



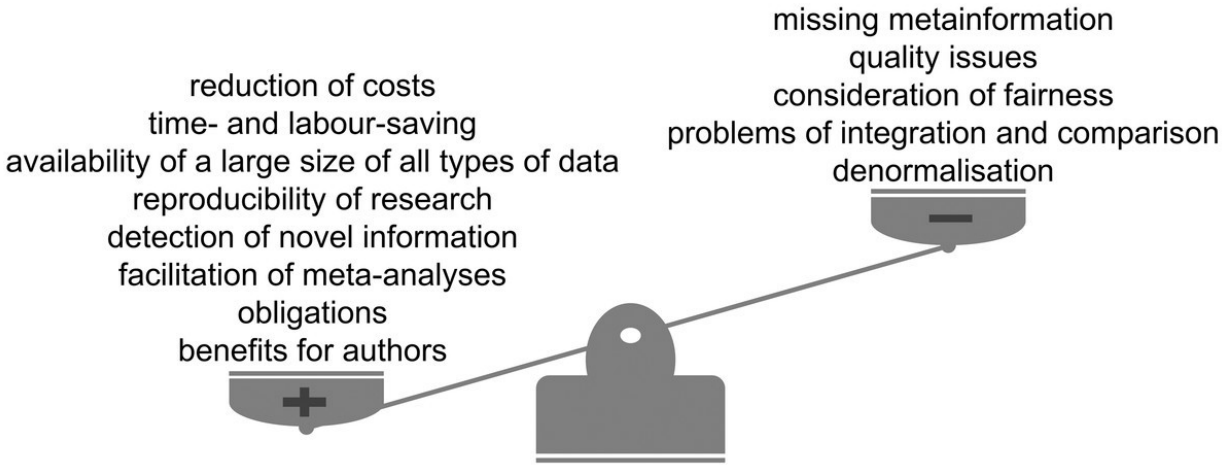
Fonte: McDonald-McGinn et al. (2011)



Fonte: Mello et al. (2018)

Balanço

- Benefícios:
  - Redução de custos
  - Redução de trabalho manual
  - Disponibilidade de diversos tipos de dados
  - Reprodutibilidade de trabalhos
  - Detecção de novas informações
  - Facilidade em meta análises
  - Benefícios aos autores.
- Precauções:
  - Perda de informação
  - Perda de qualidade
  - Uso justo de dados
  - Problemas de integração e comparação
  - Falta de normalização



Fonte: Sielemann et al. (2020)

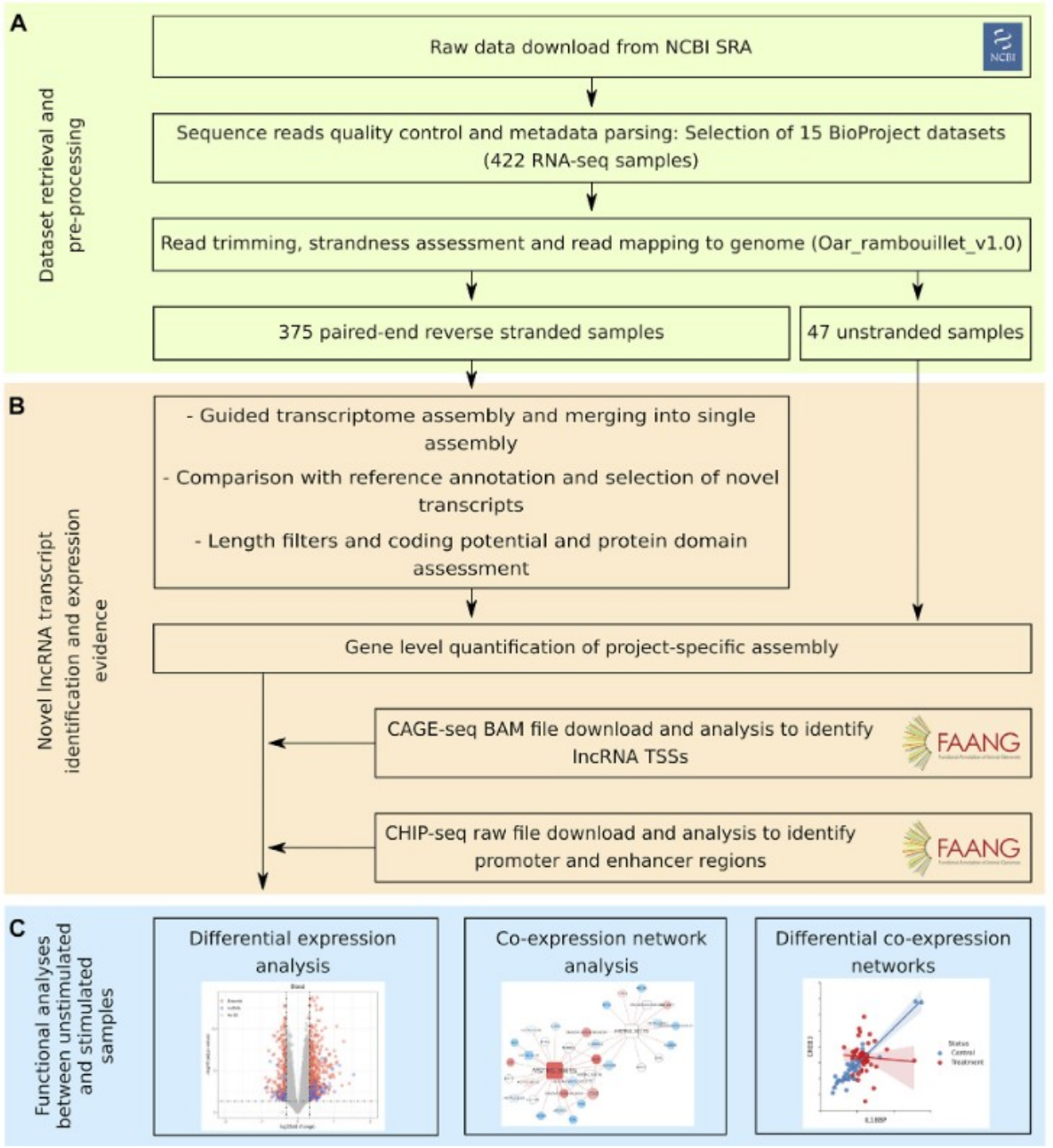
Exemplos

PMID	Title	First Author	Journal/Book	Publication Year	DOI
28480475	Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress	Meystre SM	Yearb Med Inform	2017	10.15265/IY-2017-007
26094821	Archetype-based data warehouse environment to enable the reuse of electronic health record data	Marco-Ruiz L	Int J Med Inform	2015	10.1016/j.ijmedinf.2015.05.016
30306916	Data Reuse Through Anesthesia Data Warehouse: Searching for New Use Contexts	Lamer A	Stud Health Technol Inform	2018	
29281658	Attitudes and norms affecting scientists' data reuse	Curty RG	PLoS One	2017	10.1371/journal.pone.0189288
26308551	Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide	Tenopir C	PLoS One	2015	10.1371/journal.pone.0134826
29063535	Update on Data Reuse in Health Care	Safran C	Yearb Med Inform	2017	10.15265/IY-2017-013
31164775	Credit data generators for data reuse	Pierce HH	Nature	2019	10.1038/d41586-019-01715-4
24109559	Data reuse and the open data citation advantage	Piwowar HA	PeerJ	2013	10.7717/peerj.175
30961905	A review of attitudes towards the reuse of health data among people in the European Union: The primacy of purpose and the common good	Skovgaard LL	Health Policy	2019	10.1016/j.healthpol.2019.03.012
23289463	Reuse of public genome-wide gene expression data	Rung J	Nat Rev Genet	2015	10.1038/nrg3394
31426840	Clinical trial data reuse - overcoming complexities in trial design and data sharing	Wilkinson T	Trials	2019	10.1186/s13063-019-3627-6
28002728	Expanding the Immunology Toolbox: Embracing Public-Data Reuse and Crowdsourcing	Sparks R	Immunity	2016	10.1016/j.immuni.2016.12.008
30573830	Proteomics data reuse with MassIVE-KB	Doerr A	Nat Methods	2019	10.1038/s41592-018-0283-9
23508969	CistromeFinder for ChIP-seq and DNase-seq data reuse	Sun H	Bioinformatics	2013	10.1093/bioinformatics/btt135
29881734	A Data Quality Assessment Guideline for Electronic Health Record Data Reuse	Weiskopf NG	EGEMS (Wash DC)	2017	10.5334/egems.218
26769124	SPIRIT: Systematic Planning of Intelligent Reuse of Integrated Clinical Routine Data. A Conceptual Best-practice Framework and Procedure Model	Hackl WO	Methods Inf Med	2016	10.3414/ME15-01-0045
29758095	Reuse of public, genome-wide, murine eosinophil expression data for hypotheses development	Grace JD	J Leukoc Biol	2018	10.1002/JLB.1MA1117-444R
30895574	Impact of Electronic versus Paper-Based Recording before EHR Implementation on Health Care Professionals' Perceptions of EHR Use, Data Quality, and Data Reuse	Joukes E	Appl Clin Inform	2019	10.1055/s-0039-1681054
32160189	Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide	Tenopir C	PLoS One	2020	10.1371/journal.pone.0229003
31594947	Compendiums of cancer transcriptomes for machine learning applications	Lim SB	Sci Data	2019	10.1038/s41597-019-0207-2
24395549	Drug risk assessment and data reuse	Iyasu S	Pharmacoepidemiol Drug Saf	2014	10.1002/pds.3518
26137488	Toward a Literature-Driven Definition of Big Data in Healthcare	Baro E	Biomed Res Int	2015	10.1155/2015/639021
30342679	Harmonics of Big Biomedical Data to characterize temporal variability and seasonality of data repositories: Functional Data Analysis of data temporal evolution over non-parametric statistical manifolds	Sy^ec C	Int J Med Inform	2018	10.1016/j.ijmedinf.2018.09.015
27374120	The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins	Rouillard AD	Database (Oxford)	2016	10.1093/database/baw100
29677919	From Data Extraction to Analysis: Proposal of a Methodology to Optimize Hospital Data Reuse Process	Lamer A	Stud Health Technol Inform	2018	
22733976	Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research	Weiskopf NG	J Am Med Inform Assoc	2013	10.1136/amiajnl-2011-000681
30957010	Three controversies in health data science	Peek N	Int J Data Sci Anal	2018	10.1007/s41060-018-0109-y
28906267	Evidence-Based Guidelines for Interface Design for Data Entry in Electronic Health Records	Wilbanks BA	Comput Inform Nurs	2018	10.1097/CIN.0000000000000387
29612985	Data standards can boost metabolomics research, and if there is a will there is a way	Rocca-Serra P	Metabolomics	2016	10.1007/s11306-015-0879-3
26307996	High-Performance Motion Estimation for Image Sensors with Video Compression	Xu W	Sensors (Basel)	2015	10.3390/s150820752
30439700	Modern Information Technology for Cancer Research: What's in IT for Me? An Overview of Technologies and Approaches	Ahlbrandt J	Oncology	2018	10.1159/000489368
27170913	iMacuSweb: A Tool to Support Data Reuse in Ophthalmology	Bonetto M	IEEE J Transl Eng Health Med	2016	10.1109/ITEHM.2015.2513043
29664470	Curated compendium of human transcriptional biomarker data	Golightly NP	Sci Data	2018	10.1038/sdata.2018.66
31956328	Regulating the Secondary Use of Data for Research: Arguments Against Genetic Exceptionalism	Martani A	Front Genet	2019	10.3389/fgene.2019.01254
29362097	Raising orphans from a metadata morass: A researcher's guide to re-use of public 'omics data	Bhandary P	Plant Sci	2018	10.1016/j.plantsci.2017.10.014
31314246	Research Data Stewardship for Healthcare Professionals	Jansen P	Fundamentals of Clinical Data Science	2019	10.1007/978-3-319-99713-1_4
29726424	A Clinical Data Warehouse Based on OMOP and i2b2 for Austrian Health Claims Data	Rinner C	Stud Health Technol Inform	2018	
30306898	Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features	Chazard E	Stud Health Technol Inform	2018	
29298592	Linking temporal medical records using non-protected health information data	Bonomi L	Stat Methods Med Res	2018	10.1177/0962280217698005
30231853	AgriSeqDB: an online RNA-Seq database for functional studies of agriculturally relevant plant species	Robinson AJ	BMC Plant Biol	2018	10.1186/s12870-018-1406-2
28967481	Frameworks for Evaluating Medicines in Children	Turner MA	Clin Ther	2017	10.1016/j.clinthera.2017.09.003
26140212	Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse	Soranno PA	Gigascience	2015	10.1186/s13742-015-0067-4
29796069	Barriers to the secondary use of data in critical care	Prince K	J Intensive Care Soc	2018	10.1177/1751143717741082
24534404	Building a common pediatric research terminology for accelerating child health research	Kahn MG	Pediatrics	2014	10.1542/peds.2013-1504
26267019	Neuroinformatics Software Applications Supporting Electronic Data Capture, Management, and Sharing for the Neuroimaging Community	Nichols BN	Neuropsychol Rev	2015	10.1007/s11065-015-9283-x
24625777	Refining gold from existing data	Wade TD	Curr Opin Allergy Clin Immunol	2014	10.1097/ACI.0000000000000051
31607804	Announcing the Journal of the Medical Library Association's data sharing policy	Akers KG	J Med Libr Assoc	2019	10.5195/jmla.2019.801
30612785	What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse	Looten V	Comput Methods Programs Biomed	2018	10.1016/j.cmpb.2018.12.030
29557907	DASH, the data and specimen hub of the National Institute of Child Health and Human Development	Hasra R	Sci Data	2018	10.1038/sdata.2018.46
24395550	RE: Drug risk assessment and data reuse	Toh S	Pharmacoepidemiol Drug Saf	2014	10.1002/pds.3525
30794049	Data Dissemination: Shortening the Long Tail of Traumatic Brain Injury Dark Data	Hawkins BE	J Neurotrauma	2019	10.1089/neu.2018.6192
28637268	Predicting structured metadata from unstructured metadata	Posch L	Database (Oxford)	2016	10.1093/database/baw080
29346583	DataMed - an open source discovery index for finding biomedical datasets	Chen X	J Am Med Inform Assoc	2018	10.1093/jamia/ocx121
31590214	Evaluation of Research Accessibility and Data Elements of HIV Registries	Mayer CS	Curr HIV Res	2019	10.2174/1570162X17666190924195439
32102694	Large-scale public data reuse to model immunotherapy response and resistance	Fu J	Genome Med	2020	10.1186/s13073-020-0721-z
31437900	Interactive Machine Learning for Laboratory Data Integration	Fillmore N	Stud Health Technol Inform	2019	10.3233/SHTI190198
24245661	Impacts of structuring nursing records: a systematic review	Saranto K	Scand J Caring Sci	2014	10.1111/scs.12094
27899630	European Nucleotide Archive in 2016	Toribio AL	Nucleic Acids Res	2017	10.1093/nar/gkw1106
30306905	Comparison of Changes in the Number of Included Patients Between Interventional Trials and Observational Studies Published from 1995 to 2014 in Three Leading Journals	Dezeteve A	Stud Health Technol Inform	2018	
18998885	Envisioning a biomedical data reuse registry	Piwowar HA	AMIA Annu Symp Proc	2008	
27570508	Data Citation in Neuroimaging: Proposed Best Practices for Data Identification and Attribution	Honor LB	Front Neuroinform	2016	10.3389/fninf.2016.00034
28703222	Transparency of genetic testing services for 'health, wellness and lifestyle': analysis of online prepurchase information for UK consumers	Hall JA	Eur J Hum Genet	2017	10.1038/ejhg.2017.75
24565381	Advancing data reuse in phyloinformatics using an ontology-driven Semantic Web approach	Panahiazar M	BMC Med Genomics	2013	10.1186/1755-8794-6-S3-S5
31409798	PlatformTM, a standards-based data custodianship platform for translational medicine research	Emam I	Sci Data	2019	10.1038/s41597-019-0156-9
30300549	Individual Variability of Protein Expression in Human Tissues	Kushner IK	J Proteome Res	2018	10.1021/acs.jproteome.8b00580
28881939	Data Management Rubric for Video Data in Organismal Biology	Brainerd EL	Integr Comp Biol	2017	10.1093/icb/icx060
29383564	Large Scale Analyses and Visualization of Adaptive Amino Acid Changes Projects	VV^quez N	Interdiscip Sci	2018	10.1007/s12539-018-0282-7
25415993	Big heart data: advancing health informatics through data sharing in cardiovascular imaging	Suinesiaputra A	IEEE J Biomed Health Inform	2015	10.1109/JBHI.2014.2370952
28059691	Establishing data-intensive healthcare: the case of Hospital Electronic Prescribing and Medicines Administration systems in Scotland	Cresswell K	J Innov Health Inform	2016	10.14236/jhi.v23i3.842
31483250	Designing a Framework of Components to Support Patient Engagement in Research	Vilfelein M	Stud Health Technol Inform	2019	10.3233/SHTI190800
29043940	Using Maps to Promote Data-Driven Decision-Making: One Library's Experience in Data Visualization Instruction	Catalano MM	Med Ref Serv Q	2017	10.1080/02763869.2017.1369292
31984969	A tale of three subspecialties: Diagnosis recording patterns are internally consistent but Specialty-Dependent	Diaz-Garelli JF	AMIA Open	2019	10.1093/amiaopen/ooz020
30175711	On the Way to Close the Loop in Information Logistics: Data from the Patient - Value for the Patient	Hackl WO	Yearb Med Inform	2018	10.1055/s-0038-1667076
18769183	The first step toward data reuse: disambiguating concept representation of the locally developed ICU nursing flowsheets	Kim H	Comput Inform Nurs	2008	10.1097/01.NCN.0000304839.59831.28
28455596	Coding association support systems in child and adolescent psychiatry: a systematic review	Kopsov K	Eur Child Adolesc Psychiatry	2017	10.1007/s00787-017-0992-0
30951089	An Integrated Genotyping-by-Sequencing Polymorphism Map for Over 10,000 Sorghum Genotypes	Hu Z	Plant Genome	2019	10.3835/plantgenome2018.06.0044
31315750	Constructing Large Scale Cohort for Clinical Study on Heart Failure with Electronic Health Record in Regional Healthcare Platform: Challenges and Strategies in Data Reuse	Liu DW	Chin Med Sci J	2019	10.24920/003579
30941868	Digital expression explorer 2: a repository of uniformly processed RNA sequencing data	Ziemann M	Gigascience	2019	10.1093/gigascience/giz022
22255972	Cubic spline interpolation with overlapped window and data reuse for on-line Hilbert Huang transform biomedical microprocessor	Chang NF	Conf Proc IEEE Eng Med Biol Soc	2011	10.1109/IEMBS.2011.6091792
26112029	The center for expanded data annotation and retrieval	Musen MA	J Am Med Inform Assoc	2015	10.1093/jamia/ocv048
22447219	Towards a human proteomics atlas	Gonnelli G	Anal Bioanal Chem	2012	10.1007/s00216-012-5940-8
32013947	Data Access Committees	Cheah PY	BMC Med Ethics	2020	10.1186/s12910-020-0453-z
31659323	A framework for tiered informed consent for health genomic research in Africa	Nembaware V	Nat Genet	2019	10.1038/s41588-019-0520-x
28423851	A Standardized and Data Quality Assessed Maternal-Child Care Integrated Data Repository for Research and Monitoring of Best Practices: A Pilot Project in Spain	Sy^ec C	Stud Health Technol Inform	2017	
27092246	Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources	Bousfield D	F1000Res	2016	10.12688/f1000research.7911.1
21385382	Flow cytometry data standards	Spidlen J	BMC Res Notes	2011	10.1186/1756-0500-4-50
27693565	The Learning Healthcare System: Where are we now? A systematic review	Budrionis A	J Biomed Inform	2016	10.1016/j.jbi.2016.09.018
17871727	Sharing and reusing gene expression profiling data in neuroscience	Wan X	Neuroinformatics	2007	10.1007/s12021-007-0012-5
29295106	What Information Does Your EHR Contain? Automatic Generation of a Clinical Metadata Warehouse (CMDW) to Support Identification and Data Access Within Distributed Clinical Research Networks	Bruland P	Stud Health Technol Inform	2017	
27071684	Elderly Surgical Patients: Automated Computation of Healthcare Quality Indicators by Data Reuse of EHR	Ficheur G	Stud Health Technol Inform	2016	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
28479989	A Web Based Tool to Enhance Monitoring and Retention in Care for Tuberculosis Affected Patients	Giannini B	Stud Health Technol Inform	2017	
27846806	GEN3VA: aggregation and analysis of gene expression signatures from related studies	Gundersen GW	BMC Bioinformatics	2016	10.1186/s12859-016-1321-1
23952476	Electronic health records: new opportunities for clinical research	Coorevits P	J Intern Med	2013	10.1111/joim.12119
27014709	Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies	Gini R	EGEMS (Wash DC)	2016	10.13063/2327-9214.1189
30919398	An openEHR Approach to Detailed Clinical Model Development: Tobacco Smoking Summary Archetype as a Case Study	Wei PC	Appl Clin Inform	2019	10.1055/s-0039-1681074
29192569	[From record keeping to scientific research: obstacles and opportunities for research with electronic health records]	Scholte RA	Ned Tijdschr Geneeskd	2017	
19151882	Perspectives for medical informatics. Reusing the electronic medical record for clinical research	Prokosch HU	Methods Inf Med	2009	
31390350	Temporal variability analysis reveals biases in electronic health records due to hospital process reengineering interventions over seven years	PvØrez-Benito FJ	PLoS One	2019	10.1371/journal.pone.0220369
27332243	Testing of Triggers by Data Mining of Epilepsy Patients' Structured Nursing Records	Kinnunen UM	Stud Health Technol Inform	2016	
23304327	ClinData Express--a metadata driven clinical research data management system for secondary use of clinical data	Li Z	AMIA Annu Symp Proc	2012	
28118921	Reusability of coded data in the primary care electronic medical record: A dynamic cohort study concerning cancer diagnoses	Sollie A	Int J Med Inform	2017	10.1016/j.ijmedinf.2016.08.004
23633602	Integrated database of information from structural genomics experiments	Asada Y	Acta Crystallogr D Biol Crystallogr	2013	10.1107/S0907444913001728
27529152	Quantitative monitoring of Arabidopsis thaliana growth and development using high-throughput plant phenotyping	Arend D	Sci Data	2016	10.1038/sdata.2016.55
24747879	The need for harmonized structured documentation and chances of secondary use - results of a systematic analysis with automated form comparison for prostate and breast cancer	Krumm R	J Biomed Inform	2014	10.1016/j.jbi.2014.04.008
23244533	Semantically enabling a genome-wide association study database	Beck T	J Biomed Semantics	2012	10.1186/2041-1480-3-9
24018520	Clinical professional governance for detailed clinical models	Goossen W	Stud Health Technol Inform	2013	
25848563	Recommendations for the use of operational electronic health record data in comparative effectiveness research	Hersh WR	EGEMS (Wash DC)	2013	10.13063/2327-9214.1018
25091808	Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances	Sy^ec C	Stat Methods Med Res	2017	10.1177/0962280214545122
31438173	SmartCRF: A Prototype to Visualize, Search and Annotate an Electronic Health Record from an i2b2 Clinical Data Warehouse	Cossin S	Stud Health Technol Inform	2019	10.3233/SHTI190476
32031623	The Data Tags Suite (DATS) model for discovering data access and use requirements	Alter G	Gigascience	2020	10.1093/gigascience/giz165
26152952	A Query Tool Enabling Clinicians and Researchers to Explore Patient Cohorts	Lim Choi Keung SN	Stud Health Technol Inform	2015	
28018995	Evaluating the Paper-to-Screen Translation of Participant-Aided Sociograms with High-Risk Participants	Hogan B	Proc SIGCHI Conf Hum Factor Comput Syst	2016	10.1145/2858036.2858368
29466022	From Liguria HIV Web to Liguria Infectious Diseases Network: How a Digital Platform Improved Doctors' Work and Patients' Care	Giannini B	AIDS Res Hum Retroviruses	2018	10.1089/aid.2017.0064
28625880	Predicting biomedical metadata in CEDAR: A study of Gene Expression Omnibus (GEO)	Panahiazar M	J Biomed Inform	2017	10.1016/j.jbi.2017.06.017
28049465	Secure and scalable deduplication of horizontally partitioned health data for privacy-preserving distributed statistical computation	Yigzaw KY	BMC Med Inform Decis Mak	2017	10.1186/s12911-016-0389-x
31029262	EMIF Catalogue: A collaborative platform for sharing and reusing biomedical data	Oliveira JL	Int J Med Inform	2019	10.1016/j.ijmedinf.2019.02.006
28974262	A generic method for improving the spatial interoperability of medical and ecological databases	Ghenassia A	Int J Health Geogr	2017	10.1186/s12942-017-0109-5
26306235	Rediscovering drug side effects: the impact of analytical assumptions on the detection of associations in EHR data	Diaz-Garelli JF	AMIA Jt Summits Transl Sci Proc	2015	
26538343	Discussion of "Combining Health Data Uses to Ignite Health System Learning"	Denaxas S	Methods Inf Med	2015	10.3414/ME15-12-0004
24996874	A patient-centered longitudinal care plan: vision versus reality	Dykes PC	J Am Med Inform Assoc	2014	10.1136/amiajnl-2013-002454
21262390	A comparison of two methods for retrieving ICD-9-CM data: the effect of using an ontology-based method for handling terminology changes	Yu AC	J Biomed Inform	2011	10.1016/j.jbi.2011.01.005
24302285	Data governance requirements for distributed clinical research networks: triangulating perspectives of diverse stakeholders	Kim KK	J Am Med Inform Assoc	2014	10.1136/amiajnl-2013-002308
21859004	A fast forward projection using multithreads for multirays on GPUs in medical image reconstruction	Chou CY	Med Phys	2011	10.1118/1.3591994
25065034	The promotion of data sharing in pharmacoepidemiology	Sethi N	Eur J Health Law	2014	10.1163/15718093-12341323
23201637	Harmonization and semantic annotation of data dictionaries from the Pharmacogenomics Research Network: a case study	Zhu Q	J Biomed Inform	2013	10.1016/j.jbi.2012.11.004
27071200	Embedded Streaming Deep Neural Networks Accelerator With Applications	Dundar A	IEEE Trans Neural Netw Learn Syst	2017	10.1109/TNNLS.2016.2545298
25150250	ProtocolNavigator: emulation-based software for the design, documentation and reproduction biological experiments	Khan IA	Bioinformatics	2014	10.1093/bioinformatics/btu554
27803687	Emerging Putative Associations between Non-Coding RNAs and Protein-Coding Genes in Neuropathic Pain: Added Value from Reusing Microarray Data	Raju HB	Front Neurol	2016	10.3389/fneur.2016.00168
26958300	Exploration of Temporal ICD Coding Bias Related to Acute Diabetic Conditions	McKillop M	AMIA Annu Symp Proc	2015	
32025475	The Argument Web: an Online Ecosystem of Tools, Systems and Services for Argumentation	Reed C	Philos Technol	2017	10.1007/s13347-017-0260-8
32151769	A content-based literature recommendation system for datasets to improve data reusability - A case study on Gene Expression Omnibus (GEO) datasets	Patra BG	J Biomed Inform	2020	10.1016/j.jbi.2020.103399
30122253	Earthquake prediction, biological clocks, and the cold war psy-ops: Using animals as seismic sensors in the 1970s California	Aronova E	Stud Hist Philos Sci	2018	10.1016/j.shpsa.2018.05.008
25653123	The importance of metadata to assess information content in digital reconstructions of neuronal morphology	Parekh R	Cell Tissue Res	2015	10.1007/s00441-014-2103-6
25000008	A Hybrid EAV-Relational Model for Consistent and Scalable Capture of Clinical Research Data	Khan O	Stud Health Technol Inform	2014	
31331268	JPhyloIO: a Java library for event-based reading and writing of different phylogenetic file formats through a common interface	Stöver BC	BMC Bioinformatics	2019	10.1186/s12859-019-2982-3
23793050	Knowledge management and informatics considerations for comparative effectiveness research: a case-driven exploration	Embi PJ	Med Care	2013	10.1097/MLR.0b013e31829b1de1
27376095	Performance of a Natural Language Processing (NLP) Tool to Extract Pulmonary Function Test (PFT) Reports from Structured and Semistructured Veteran Affairs (VA) Data	Sauer BC	EGEMS (Wash DC)	2016	10.13063/2327-9214.1217
31959844	Ontology-driven integrative analysis of omics data through Onassis	Galeota E	Sci Rep	2020	10.1038/s41598-020-57716-1
31424314	Higher body mass index is not a protective risk factor for 28-days mortality in critically ill patients with acute kidney injury undergoing continuous renal replacement therapy	Wang H	Ren Fail	2019	10.1080/0886022X.2019.1650767
18661966	SPECTRa: the deposition and validation of primary chemistry research data in digital repositories	Downing J	J Chem Inf Model	2008	10.1021/ci700473?
30595024	Evolution of Dementia Related to the Use of Alcohol in the French Nationwide Discharge Summary Database Between 2007 and 2017	Rochoy M	Am J Alzheimers Dis Other Demen	2019	10.1177/1533317518822043
24303283	Using clinical element models for pharmacogenomic study data standardization	Zhu Q	AMIA Jt Summits Transl Sci Proc	2013	
26998241	ISCB's initial reaction to New England Journal of Medicine editorial on data sharing	Valencia A	F1000Res	2016	10.12688/f1000research.8051.1
28876906	Water Issues Related to Transitioning from Conventional to Unconventional Oil Production in the Permian Basin	Scanlon BR	Environ Sci Technol	2017	10.1021/acs.est.7b02185
25463966	Using electronic health records for clinical research: the case of the EHR4CR project	De Moor G	J Biomed Inform	2015	10.1016/j.jbi.2014.10.006
24126951	[Implementation of a patient data management system. Effects on intensive care documentation]	Castellanos I	Anaesthesist	2013	10.1007/s00101-013-2239-x
14625855	Hierarchical analysis of large-scale two-dimensional gel electrophoresis experiments	Rubinfeld A	Proteomics	2003	10.1002/pmic.200300533
25848610	Sustainability considerations for health research and analytic data infrastructures	Wilcox A	EGEMS (Wash DC)	2014	10.13063/2327-9214.1113
27444186	[Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes	Chiaramello E	J Biomed Inform	2016	10.1016/j.jbi.2016.07.017
30756117	Shift in Hospitalizations for Alzheimer's Disease to Related Dementias in France between 2007 and 2017	Rochoy M	J Prev Alzheimers Dis	2019	10.14283/jpad.2019.5
10144371	Information technology in the foxhole	Eyestone SM	Top Health Inf Manage	1995	
32171029	Enabling reusability of plant phenomic datasets with MIAPPE 1.1	Papoutsoglou EA	New Phytol	2020	10.1111/nph.16544
31040006	Open access image repositories: high-quality data to enable machine learning research	Prior F	Clin Radiol	2020	10.1016/j.crad.2019.04.002
21342582	FPGA accelerator for protein secondary structure prediction based on the GOR algorithm	Xia F	BMC Bioinformatics	2011	10.1186/1471-2105-12-S1-S5
21678495	AnatomicalTerms.info: heading for an online solution to the anatomical synonym problem hurdles in data-reuse from the Terminologia Anatomica and the foundational model of anatomy and potentials for future development	GobvÈe OP	Clin Anat	2011	10.1002/ca.21185
28466088	Combining Contrast Mining with Logistic Regression To Predict Healthcare Utilization in a Managed Care Population	Sheets L	Appl Clin Inform	2017	10.4338/ACI-2016-05-RA-0078
21893862	Recording associated disorders using SNOMED CT	Cornet R	Stud Health Technol Inform	2011	
20442144	International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the next generation HIPAA code sets	Steindel SJ	J Am Med Inform Assoc	2010	10.1136/jamia.2009.001230
32324446	Impact on Quality of Documentation and Workload of the Introduction of a National Information Standard for Tumor Board Reporting	Ebben KCWJ	JCO Clin Cancer Inform	2020	10.1200/CCI.19.00050
25428570	Empirical study using network of semantically related associations in bridging the knowledge gap	Abedi V	J Transl Med	2014	10.1186/s12967-014-0324-9
25967024	Region-referenced phase unwrapping architecture for digital holographic microscopy	Hwang WJ	Appl Opt	2015	10.1364/AO.54.004467
32308882	De-identification of Clinical Text via Bi-LSTM-CRF with Neural Language Models	Tang B	AMIA Annu Symp Proc	2020	
25848591	Learning from health information exchange technical architecture and implementation in seven beacon communities	McCarthy DB	EGEMS (Wash DC)	2014	10.13063/2327-9214.1060
32301862	Research on the Radiotoxicology of Plutonium Using Animals: Consideration of the 3Rs-Replace, Reduce, Refine	Griffiths NM	Health Phys	2020	10.1097/HP.0000000000001258
25848582	Validation of Diagnostic and Procedural Codes for Identification of Acute Cardiovascular Events in US Veterans with Rheumatoid Arthritis	Davis LA	EGEMS (Wash DC)	2014	10.13063/2327-9214.1023
32521555	Factors Influencing Problem List Use in Electronic Health Records-Application of the Unified Theory of Acceptance and Use of Technology	Klappe ES	Appl Clin Inform	2020	10.1055/s-0040-1712466
21346966	Development of automated quality reporting: aligning local efforts with national standards	Dykes PC	AMIA Annu Symp Proc	2010	
24808228	Efficient acceleration of mutual information computation for nonrigid registration using CUDA	Ikeda K	IEEE J Biomed Health Inform	2014	10.1109/JBHI.2014.2310745
32418147	Development, implementation and preliminary evaluation of clinical dashboards in a department of anesthesia	Laurenti G	J Clin Monit Comput	2020	10.1007/s10877-020-00522-x
19208138	Fine-grained parallel RNAalifold algorithm for RNA secondary structure prediction on FPGA	Xia F	BMC Bioinformatics	2009	10.1186/1471-2105-10-S1-S37

Fonte: Sielemann et al. (2020)

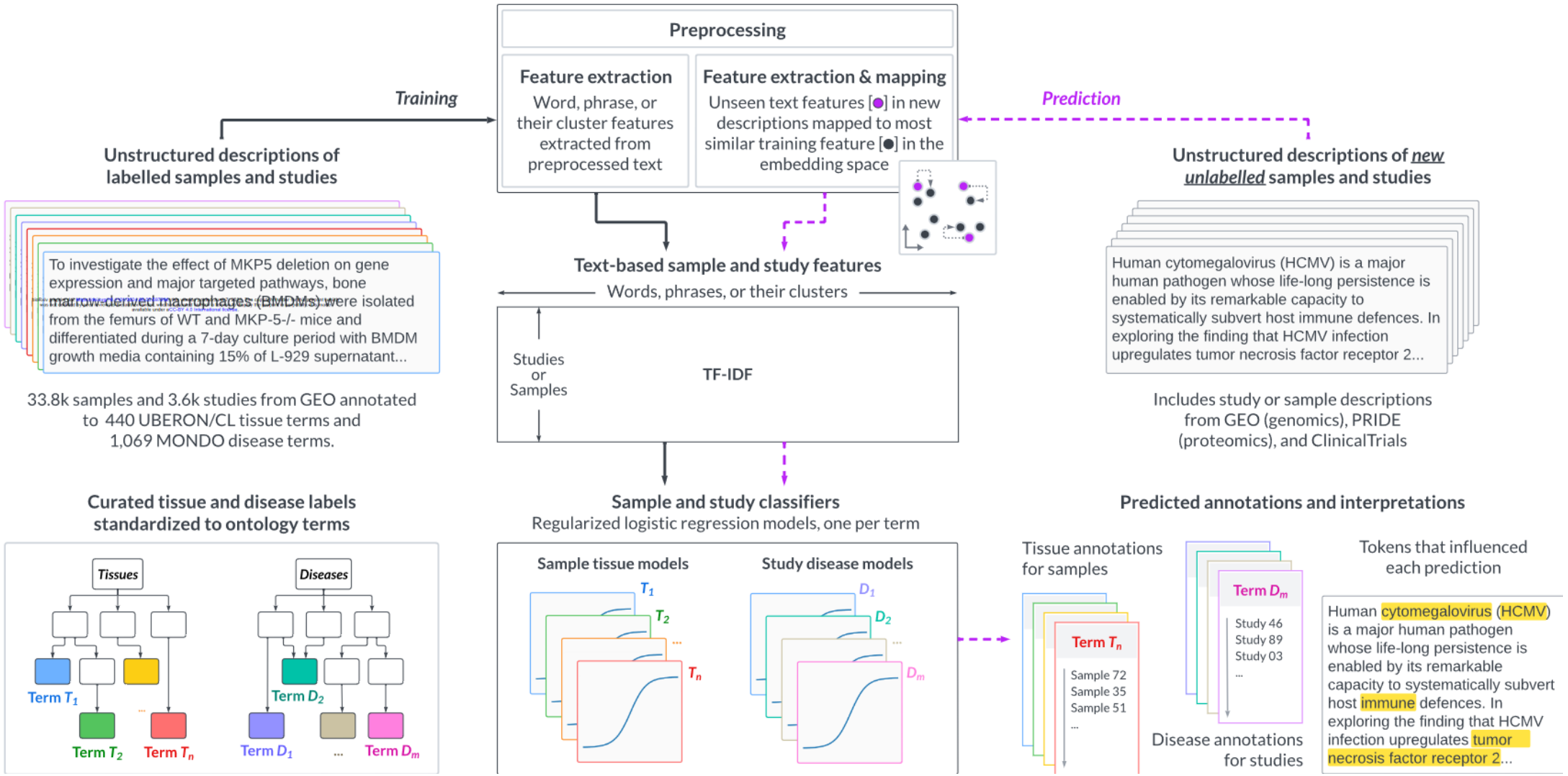
Exemplos

- Utilização de diversos conjuntos de dados de amostras do sistema immune de ovinos
- Identificação de lncRNAs associados à resposta immune em ovinos



Fonte: Bilbao-Arribas e Jugo (2022)

Exemplos



Fonte: Yuan et al. (2024)

Referências

- Arend D, Lange M, Chen J, Colmsee C, Flemming S, Hecht D, Scholz U. 2014. e!DAL - a framework to store, share and publish research data. BMC Bioinformatics. 15(1):214.
- Bierer BE, Crosas M, Pierce HH. 2017. Data Authorship as an Incentive to Data Sharing. New England Journal of Medicine. 376(17):1684–1687.
- Bilbao-Arribas M, Jugo BM. 2022. Transcriptomic meta-analysis reveals unannotated long non-coding RNAs related to the immune response in sheep. Front Genet. 13. [accessed 2024 Jul 3]. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.1067350/full.
- Borghi J, Van Gulick A. 2022. Promoting Open Science Through Research Data Management. Harvard Data Science Review. [accessed 2024 Jul 3]. https://hdsr.mitpress.mit.edu/pub/72kcw990.
- Cunha-Oliveira T, Ioannidis JPA, Oliveira PJ. 2024. Best practices for data management and sharing in experimental biomedical research. Physiological Reviews. 104(3):1387–1408.
- Gomes DGE, Pottier P, Crystal-Ornelas R, Hudgins EJ, Foroughirad V, Sánchez-Reyes LL, Turba R, Martinez PA, Moreau D, Bertram MG, et al. 2022. Why don’t we share data and code? Perceived barriers and benefits to public archiving practices. Proceedings of the Royal Society B: Biological Sciences. 289(1987):20221113.
- Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, Stefano RD, Gil Y, Groth P, Hedstrom M, et al. 2014. Ten Simple Rules for the Care and Feeding of Scientific Data. PLOS Computational Biology. 10(4):e1003542.
- Hafner A, DeLeo V, Deng C, Elsik CG, Fleming D, Harrison PW, Kalbfleisch TS, Petry B, Pucker B, Quezada-Rodríguez EH, et al. 2024. Data Reuse in Agricultural Genomics Research: Present Challenges and Future Solutions. [accessed 2024 Jul 3]. https://www.preprints.org/manuscript/202401.0780/v1.
- McDonald-McGinn DM, Sullivan KE. 2011. Chromosome 22q11.2 Deletion Syndrome (DiGeorge Syndrome/Velocardiofacial Syndrome). Medicine. 90(1):1.
- Mello MM, Lieou V, Goodman SN. 2018. Clinical Trial Participants’ Views of the Risks and Benefits of Data Sharing. New England Journal of Medicine. 378(23):2202–2211.
- Raman A. 2021. Secondary Data Analysis: Lessons and perspective of a research parasite. [accessed 2024 Jul 23]. https://osf.io/bec69.
- Schwab S, Janiaud P, Dayan M, Amrhein V, Panczak R, Palagi PM, Hemkens LG, Ramon M, Rothen N, Senn S, et al. 2022. Ten simple rules for good research practice. PLOS Computational Biology. 18(6):e1010139.
- Sielemann K, Hafner A, Pucker B. 2020. The reuse of public datasets in the life sciences: potential risks and rewards. PeerJ. 8:e9954.
- Yuan H, Hicks P, Ahmadian M, Johnson K, Valtadoros L, Krishnan A. 2024. Annotating publicly-available samples and studies using interpretable modeling of unstructured metadata. [accessed 2024 Jul 3]. http://biorxiv.org/lookup/doi/10.1101/2024.06.03.597206.

Obrigado!

carvalhopc@usp.br
linktr.ee/carvalhopc