



• GHGA Webinar

FAIR Data for FAIR Biomedical Portals

In cooperation with





Part 1: FAIR Biomedical Data

- Q&A Session

Part 2: FAIR Data Portals

- Q&A Session and Feedback Poll

In cooperation with



FAIR Biomedical Data

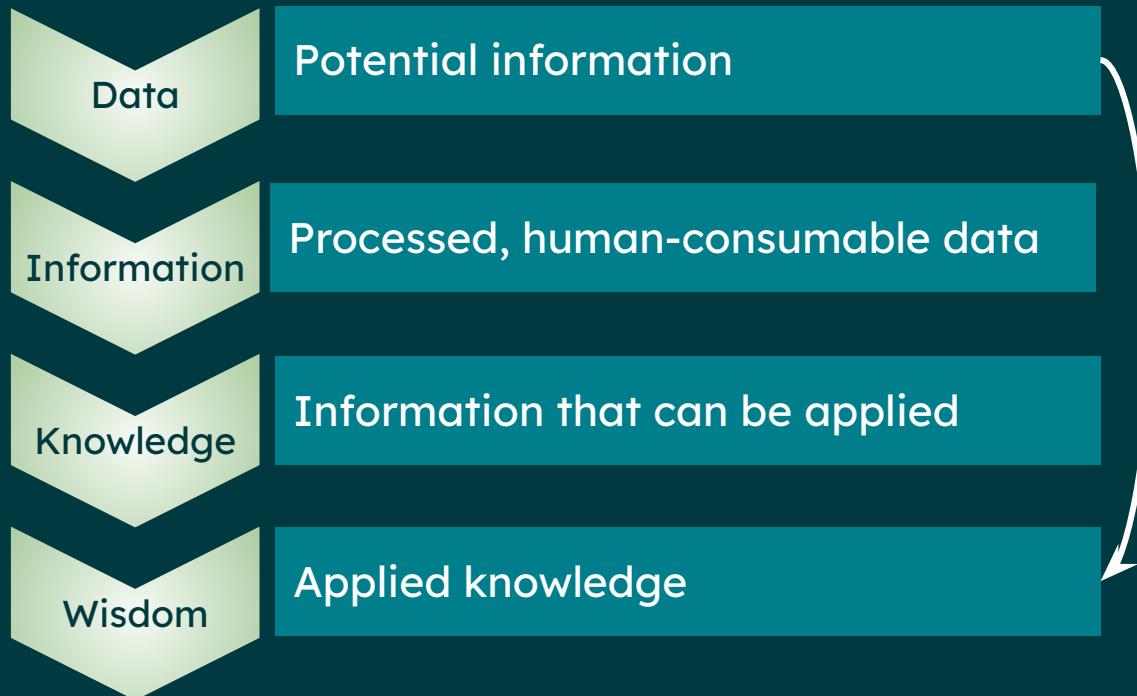
Structure

- Data and Metadata
- Biomedical Metadata
- What is FAIR...
- ...and why is it important?
- How can I make my data FAIR?
- What is a metadata model?
- Existing metadata solutions
- How to create a metadata model
- Pitfalls and things to consider

Data and Metadata

FAIR Biomedical Data

Data and Metadata



→ Extraction of information with processing and contextualizing
→ Description of the data = **METADATA**

Data and Metadata



Many FASTQ files



```
@SEQ_ID  
GATTGGGGTCAAACAGTATCGATCAAATAGTAATCC  
+  
!'''*((((*++))%++)(%%%),1***-+*'')**
```

Normalized and annotated gene counts;
differentially expressed genes



Ensembl
version
Patient
groups



Do my findings align with the
literature? What is new?



Discussion

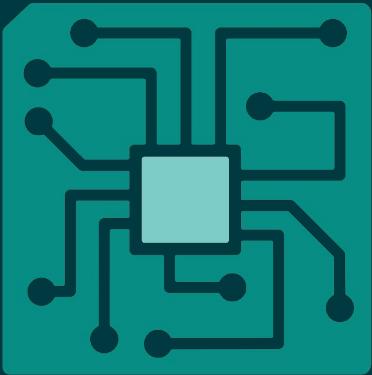
Possible treatments based on
identified genes



Follow up
studies

Replication

Biomedical metadata



GHGA

INDIVIDUAL DATA

Demographics
Clinical information
(diagnoses, phenotype, measurements)

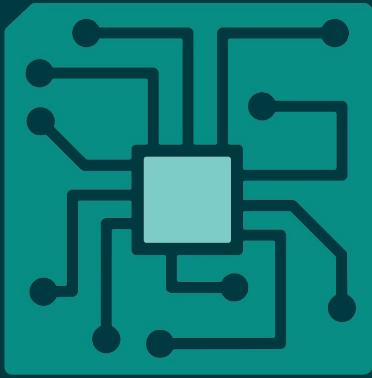
TECHNICAL DATA

Data type
Instrument model
Pipelines

“ADMINISTRATIVE” DATA

Why
Where
Who
What

Biomedical metadata: Core fields



INDIVIDUAL DATA

- age
- sex
- disease
- case / control
- sample tissue

TECHNICAL DATA

- data type
- instrument
- model
- pipelines
- software
- versions

“ADMINISTRATIVE” DATA

- study type
- data controller
- place of collection
- data modalities
- abstract

Why be FAIR

FAIR Biomedical Data

What is FAIR...



Findable



Accessible



Interoperable



Reusable

Others can find your data

Others (know how they) can access
your data

Your data is readable for humans and
machines and can be interacted with

Your data is well annotated and can
be used for diverse use cases

What is FAIR...

The **FAIR Guiding Principles for scientific data management and stewardship**
(Wilkinson et al., 2016)

„(...) act as a **guideline** for those wishing to enhance the reusability of their data holdings.“

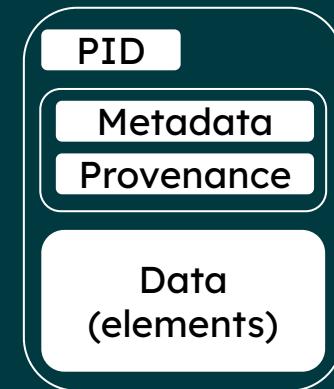
„(...) the FAIR principles put specific emphasis on enhancing the ability of **machines** to automatically find and use the data, in addition to supporting its reuse by individuals.“

Why?

- Explosion of data
 - Increased international collaboration
- Data management issues

A FAIR Digital Object

is only FAIR if all components are FAIR



... and why is it important?

Why would a researcher even want to share data?

- It is part of the scientific process
 - Review validity of findings
 - Allow others to answer further questions
- Lead by example → encourage others to share data and benefit from data availability

FAIR Data advances science and builds trust

... and why is it important?

What happens when data is not shared: The Reproducibility Crisis



Authors either do not share their data at all or make it difficult to understand; on purpose or not

„among studies that provided **SUFFICIENT** information to be redone, the effect sizes were **85% smaller on average** than the original findings“

“**original positive results** were **half as likely to replicate successfully** (40%) than original null results (80%)”



... and why is it important?

No raw data, no science: another possible source of the reproducibility crisis

Tsuyoshi Miyakawa 

Molecular Brain 13, Article number: 24 (2020) | [Cite this article](#)

66k Accesses | 122 Citations | 2082 Altmetric | [Metrics](#)

“[...] more than 97% of the 41 manuscripts did not present the raw data supporting their results when requested by an editor, suggesting a possibility that the raw data did not exist from the beginning, at least in some portions of these cases.”

<https://molecularbrain.biomedcentral.com/articles/10.1186/s13041-020-0552-2>

„[...] just getting those data sets online will not bring anticipated benefits: few data sets will really be FAIR, because most will be unfindable. What's needed are policies and infrastructure to organize metadata.”

<https://www.nature.com/articles/d41586-022-02820-7>

WORLD VIEW | 05 September 2022

Without appropriate metadata, data-sharing mandates are pointless



Funders and investigators must demand appropriate metadata standards to take data from foul to FAIR.

By [Mark A. Musen](#) 

How can I make my data FAIR?



Findable



Accessible



Interoperable



RReusable



Depends on
infrastructure and
data subjects



Depends on researcher
and implemented
standards

How can I make my data FAir?

Findable Data

→ Others can find your data

- Is the data allowed to be findable (e.g. detailed clinical data, social data)
- Findability depends on solutions implemented by a data repository
 - Search engine
 - Indexing
 - Keywords
 - Vocabulary wrangling

Accessible Data

→ Others can access your data

- Is the data allowed to be accessible or do restrictions apply
- Technical accessibility depends on solutions implemented by data repository
 - Open protocols (html)
 - (Meta)data lifecycle

How can I make my data faIR?

Interoperable Data

→ Your data is machine and human readable and can be interacted with

- Vocabulary harmonization and standardization
(ontologies, controlled vocabularies)
- Data model to describe and structure data (phenopackets, OMOP, FHIR)
- Data dictionaries explain the data and metadata

Reusable Data

→ Your data can be used for other studies and use cases

- Well annotated
- Detailed metadata
- Usage of domain-relevant community standards (fastq, hd5, csv)
- Linked to a publication with information about methods, results and limitations of the data

How can I make my data FAIR? (TL;DR)

Things you can do

- Think about what your data could deliver and make a “metadata wishlist”
- Vocabulary harmonization and standardization
- Data model to structure data
- Data dictionary
- Domain-relevant standards
- Publish methods, results and limitations of the data
- Choose your data portal wisely

Things you have to accept

- Detailed annotation might not be possible
(e.g. controlled study vs real world data)
- Standards have not been accepted in every research field
(especially upcoming fields)
- Accessibility and findability might be limited by GDPR / ethics

How can I make my data FAIR?

- Record plans and decisions on how to make your data FAIR in a data management plan

<https://ds-wizard.org/>

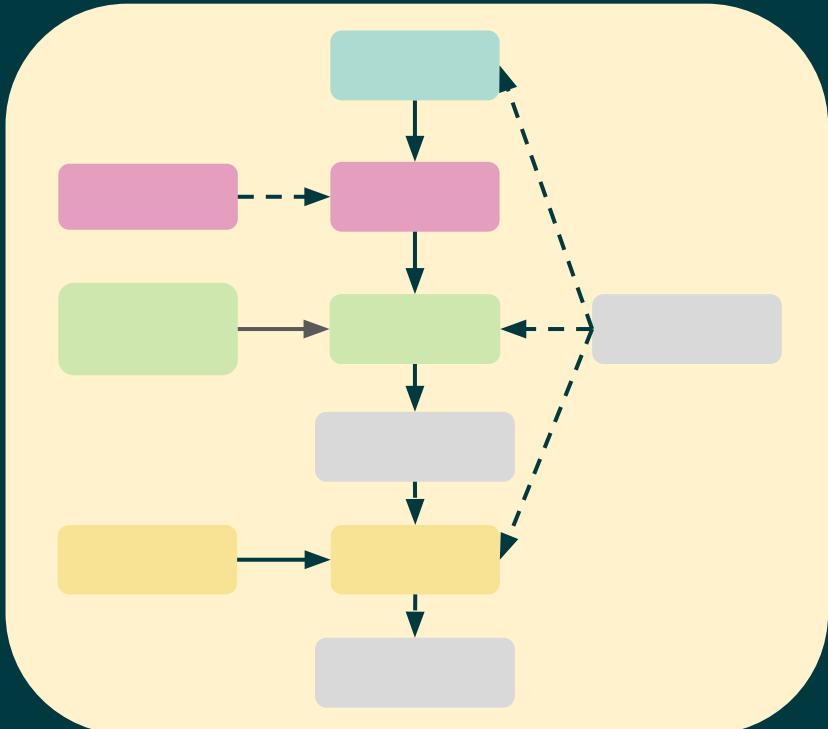
<https://rdmkit.elixir-europe.org/>



Metadata Models

FAIR Biomedical Data

What is a metadata model?



- Representation of captured data
- Allows to
 - Structure data
 - Define relationships between entities (real world objects)
 - Describe entities using properties
(entity “individual” has properties “sex”, “age”, “diagnosis”)
- Store and query data in a database

Existing metadata solutions

Agnostic solutions

- RO-Crate
- schemas.org
- LinkML
- JSON-LD
- RDF

Provenance data

- Workflow RO-Crate
- BioCompute Object

Biomedical data models

- ISA-tab
- bioschemas.org
- OMOP
- FHIR
- phenopackets

- Kerndatensatz MII
- NFDI4Health

How to create a metadata model

Identify your data

Are you dealing with a certain disease? Are you working with cohort data or real world data?

Identify necessary compatibilities

Will you always upload your data to the same portal? Does your group already use XML over JSON?

Identify already existing standards

Are there community standards you can adhere to, e.g. for your experiment or study type?

Identify entities

Which real world objects (e.g. persons, protocols) will you capture information about?

Link your entities

Linking is required to store and retrieve data. Shuffle entities around if needed.

Standardize your vocabulary

Nothing is worse than having only free text entries. Check OLS* to find suitable ontologies or define your own controlled vocabularies

Identify properties

What information do you need to describe your entities?

*OLS: Ontology Lookup Service managed by EMBL-EBI:
<https://www.ebi.ac.uk/ols4/>

Pitfalls and things to consider

Aligning to standards and other schemas is your way to go

Make use of community-accepted standards to FAIRify your data. Only develop something new if you really have to!

Know your GDPR and your DPO

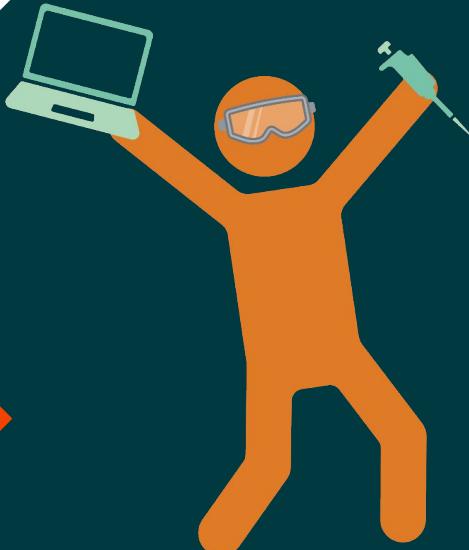
Human biomedical data is considered sensitive data under Art. 9 par. 1 GDPR and requires special protection and processing contracts.

Ontologies and controlled vocabularies are your friends

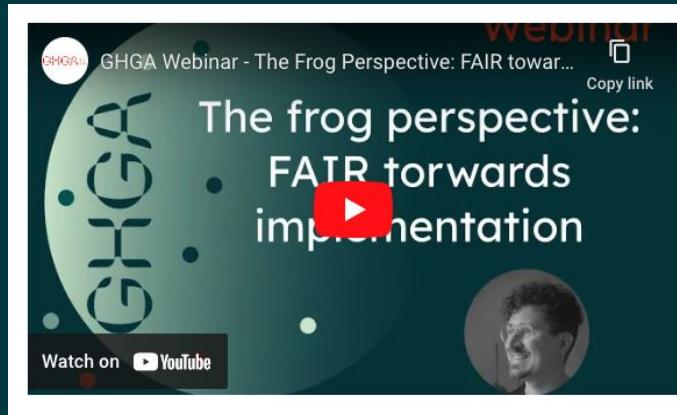
Why develop a metadata model if you're not going to control what people put in there.

Accessible data ≠ open data

To be FAIR-compliant, your data has to be accessible, not open. Just because data is open, does not mean it is accessible.



Additional resources



Communities and Data Portals

ELIXIR

- Europe's national bioinformatics nodes into a single infrastructure
- Underpins the management and safeguarding of data generated by publicly-funded research.
- Purpose:
 - Find
 - Analyse
 - Share data
 - Exchange expertise
 - Implement best practices

Distributed Europe Research Infrastructure for Life Science Data



Supporting data repositories to be FAIR

Standards	Identifier Services	FAIR Services
 Domain Standards  Bioschemas.org Schema.org universal machine processable metadata mark-up	 Identifier management, resolution & best practice  BridgeDB Identifier mapping  OLS  OXO  Omics DI	 FAIRsharing.org FAIR Repository Recommendations & FAIR indicators  FAIR cookbook FAIRification
		 Text mining AI/ML Data citation tracking Impact Analysis

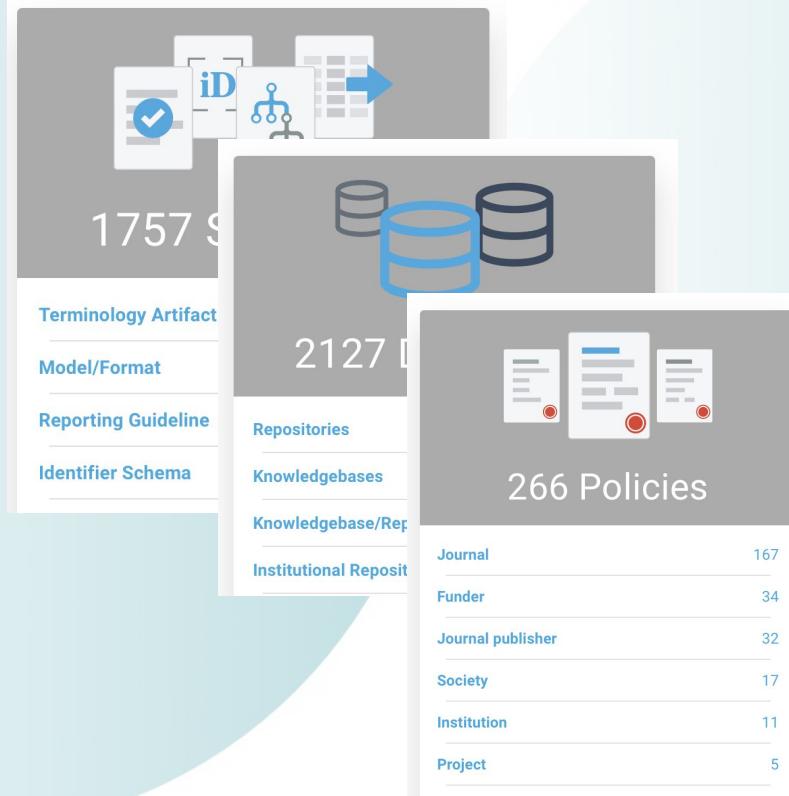
Research Data Alliance

- RDA is an international organization with a focus on infrastructure and community development that reduce barriers to data sharing and exchange
- Main focus groups:
 - Reproducibility
 - Data preservation
 - Best practices for domain repositories
 - Curriculum development
 - Data citation
 - Data type registries
 - Metadata



FAIRsharing.org

- A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies
- Open standards: reporting guidelines, terminology, and exchange formats



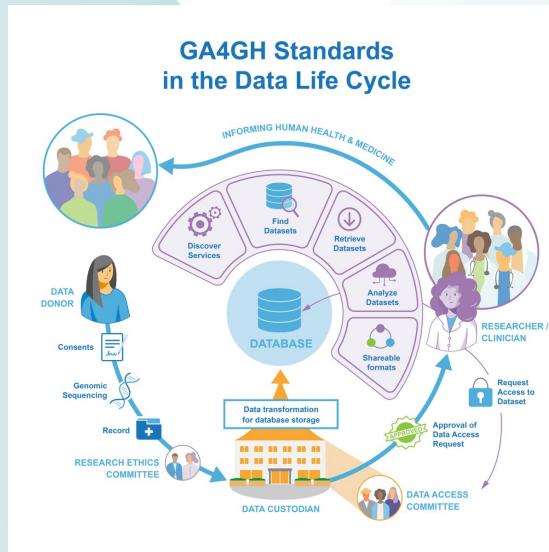
Global Alliance for Genomics and Health

- GA4GH aims to accelerate progress in genomic science and human health through

- **developing** standards
- **framing** policies

for responsible genomic and health data sharing

- Core Objectives
 - Data sharing
 - Interoperability
 - Privacy and security
 - Ethical and legal considerations
 - Innovation

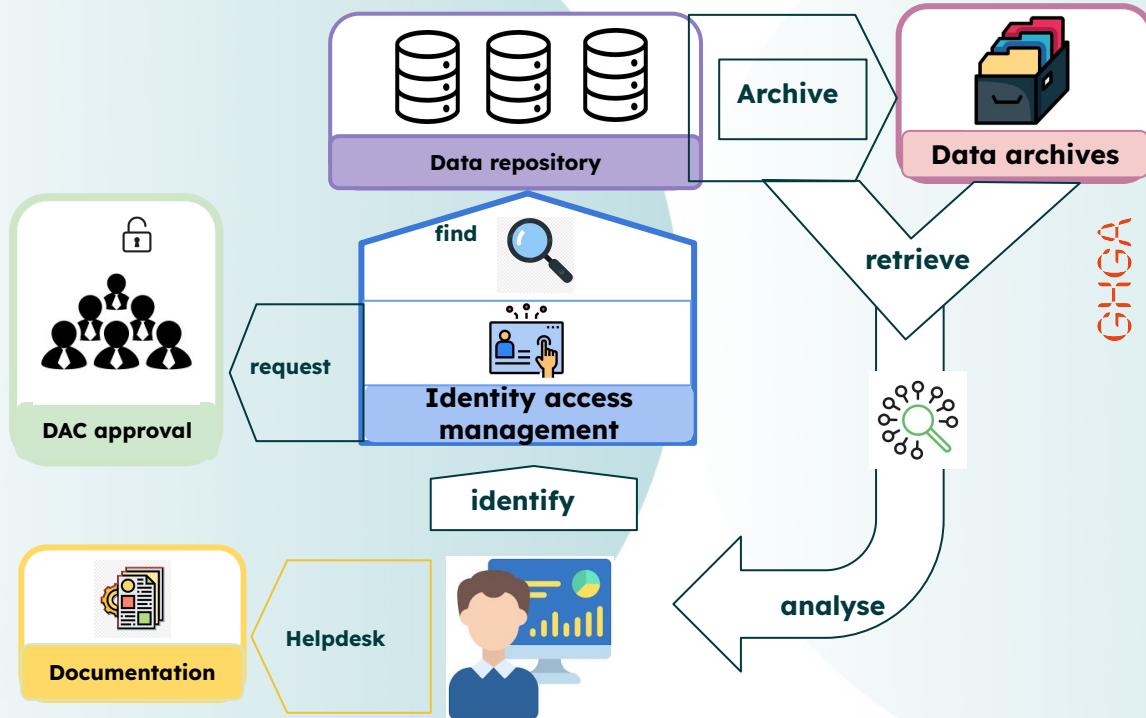


<https://www.sciencedirect.com/science/article/pii/S2666979X21000367>

Need for FAIR data portals

- Growing demand for quality research datasets that are
 - stored
 - managed
 - shared

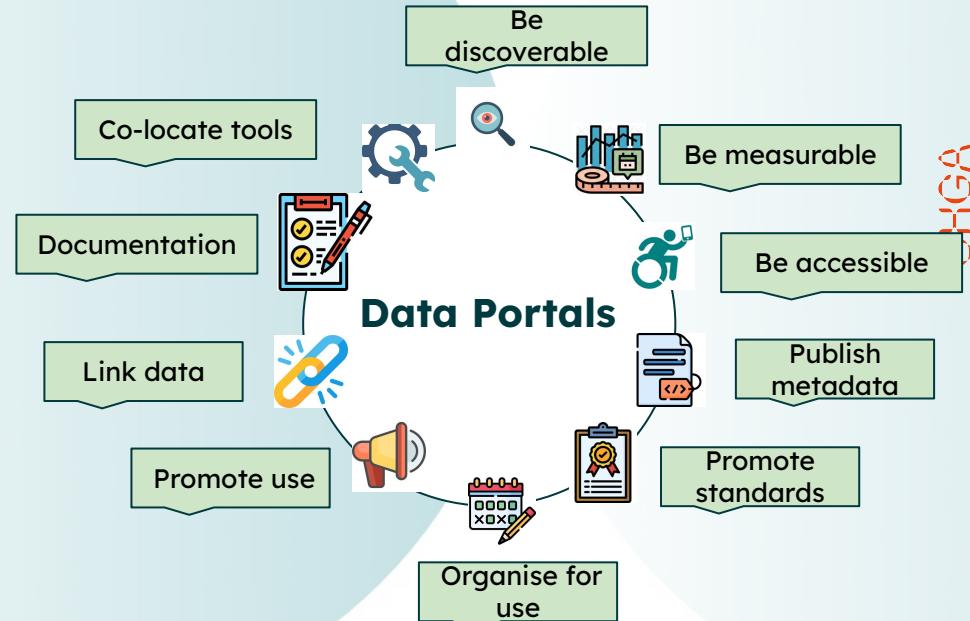
in a **reliable** and **trustworthy** way



Open Data portals

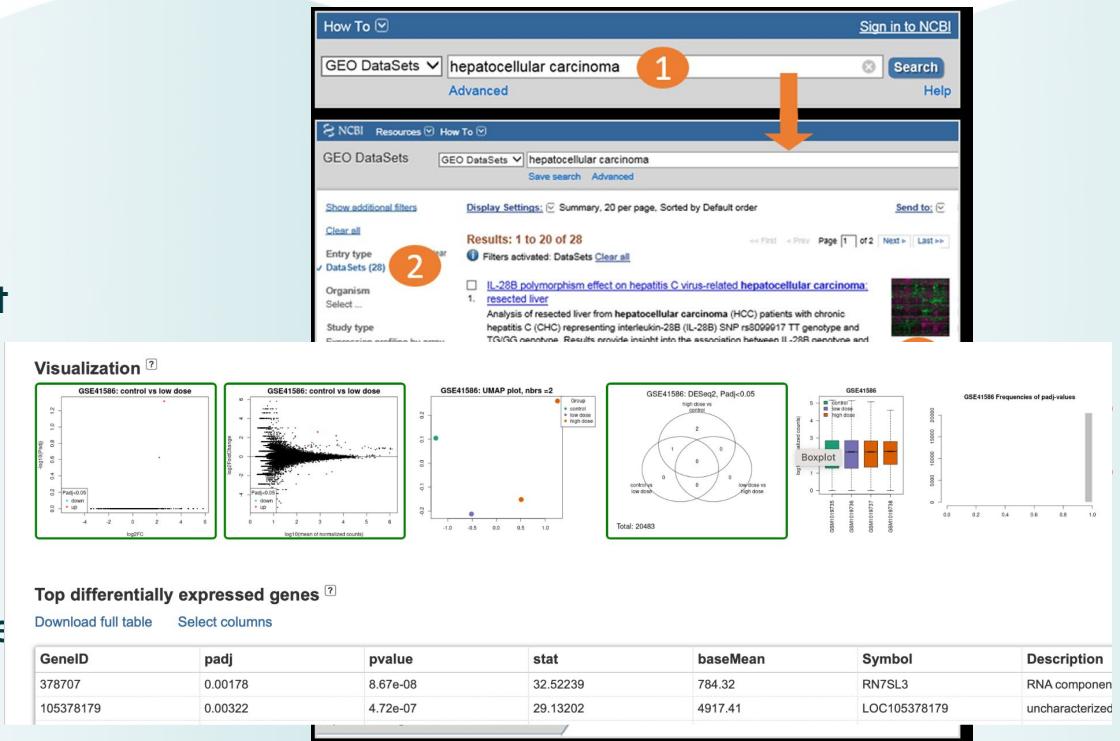
- Data portal is a platform designed to store, publish and share data from different applications and users.

- Easy to submit
- Searchable
- Transparent



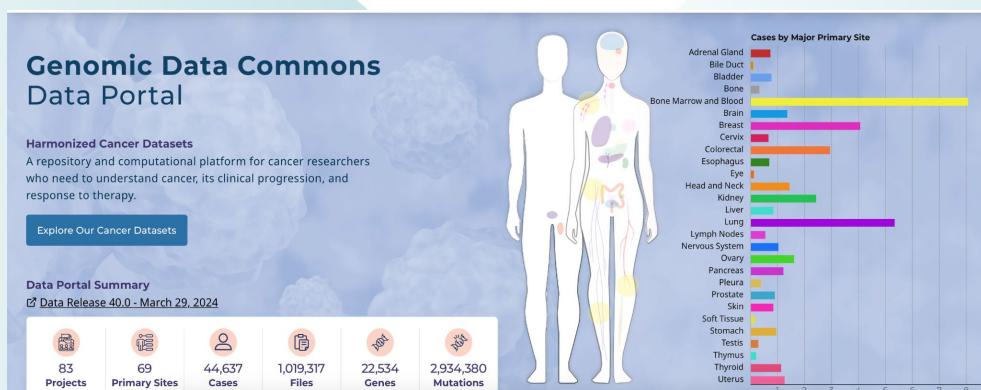
NCBI GEO

- Gene Expression Omnibus (GEO) is a database repository of high throughput gene expression data and hybridization arrays, chips, microarrays.
- Public archive and resource for gene expression data.



Genomic Data Commons

- GDC is a web-based platform that allows users to search, analyze, and download data from cancer genomic studies



- The cohort builder allows to filter datasets based on clinical, biospecimen, and available data elements to create custom cohorts.

General	
Demographic	
General Diagnosis	
Disease Status and History	
Stage Classification	16 more
Grade Classification	
Other Classification	
Treatment	
Exposure	
Biospecimen	

Program	
Name	Cases
<input type="checkbox"/> APOLLO	87 (0.19%)
<input type="checkbox"/> BEATAML1.0	882 (1.98%)
<input type="checkbox"/> CDDP_EAGLE	50 (0.11%)
<input type="checkbox"/> CGCI	645 (1.44%)
<input type="checkbox"/> CMT	299 (0.67%)
<input type="checkbox"/> CPTAC	1,656 (3.71%)
16 more	

Project	
Name	Cases
<input type="checkbox"/> APOLLO-LUAD	87 (0.19%)
<input type="checkbox"/> BEATAML1.0-COHORT	826 (1.85%)
<input type="checkbox"/> BEATAML1.0-CRENO...	56 (0.13%)
<input type="checkbox"/> CDDP_EAGLE-1	50 (0.11%)
<input type="checkbox"/> CGCI-BLGP	324 (0.73%)
<input type="checkbox"/> CGCI-HTMCP-CC	212 (0.47%)
77 more	

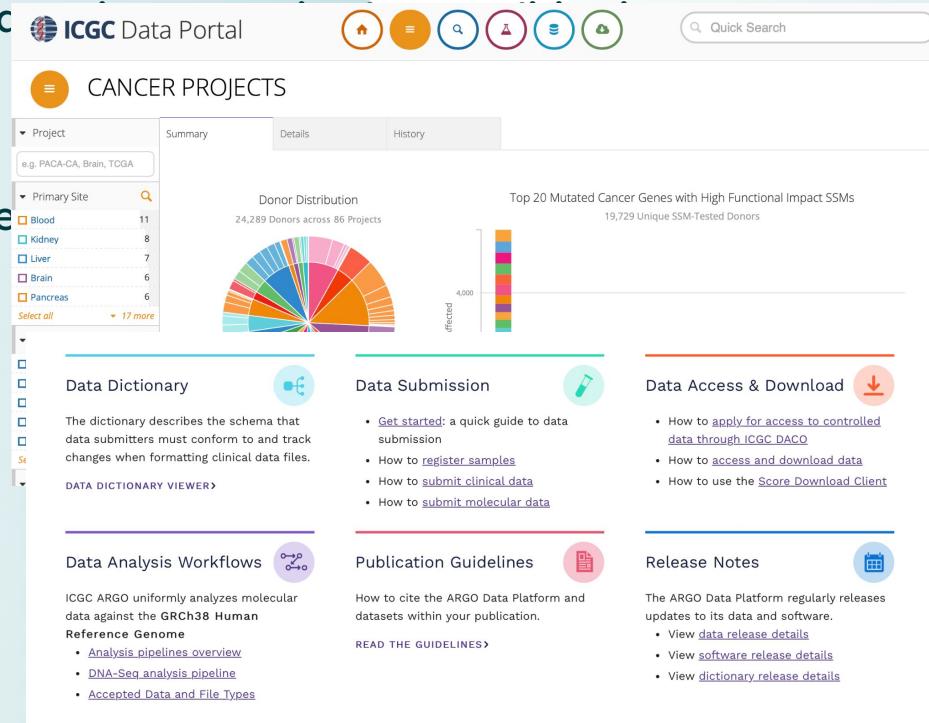
Disease Type	
Name	Cases
<input type="checkbox"/> acinar cell neoplasms	248 (0.56%)
<input type="checkbox"/> acute lymphoblastic...	1,086 (2.43%)
<input type="checkbox"/> adenomas and ade...	14,503 (32.49%)
<input type="checkbox"/> adnexal and skin ap...	22 (0.05%)
<input type="checkbox"/> basal cell neoplas...	21 (0.05%)
<input type="checkbox"/> blood vessel tumors	1 (0.00%)
39 more	

Primary Site	
Name	Cases
<input type="checkbox"/> accessory sinuses	1 (0.00%)
<input type="checkbox"/> adrenal gland	721 (1.62%)
<input type="checkbox"/> anus and anal canal	79 (0.18%)
<input type="checkbox"/> base of tongue	26 (0.06%)
<input type="checkbox"/> bladder	763 (1.71%)
<input type="checkbox"/> bones, joints and ar...	257 (0.58%)

Tissue or Organ of Origin	
Name	Cases
<input type="checkbox"/> abdomen, nos	186 (0.42%)
<input type="checkbox"/> adrenal gland, nos	568 (1.27%)
<input type="checkbox"/> ampulla of vater	4 (0.01%)
<input type="checkbox"/> anal canal	1 (0.00%)
<input type="checkbox"/> anterior floor of mo...	2 (0.00%)
<input type="checkbox"/> anterior mediastinu...	29 (0.06%)

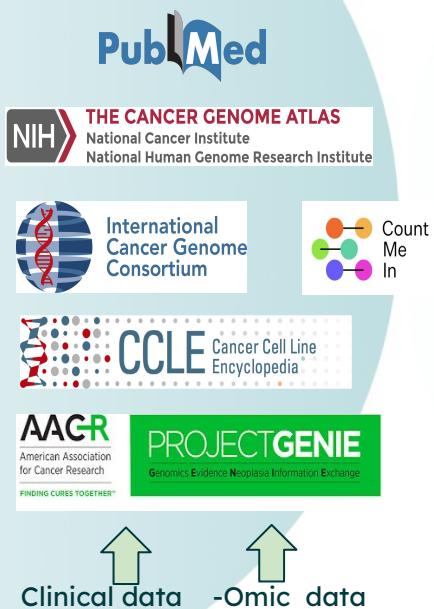
International Cancer Genome Consortium

- ICGC is a collaborative effort to characterize cancer types
- The data portal uses GUI to offer researchers analyze data.



cBioPortal

- Platform for **exploratory** and **interactive visualization, analysis** of large-scale **cancer genomics** data sets



Biological data (eg: Networks,
3D protein structure)



Curated effects and therapy
implications



GHGCA

Predicted functional effect



mutationassessor.org
functional impact of protein mutations

Variant recurrence



Ways to assess FAIRness in data portals

Implications of poor data quality

- Increasing number of research discoveries are made through collaboration and using other's data
- Impact of poor data quality
 - Inaccuracy
 - Redundancy
 - Incompleteness
 - Expensive



Garbage in!



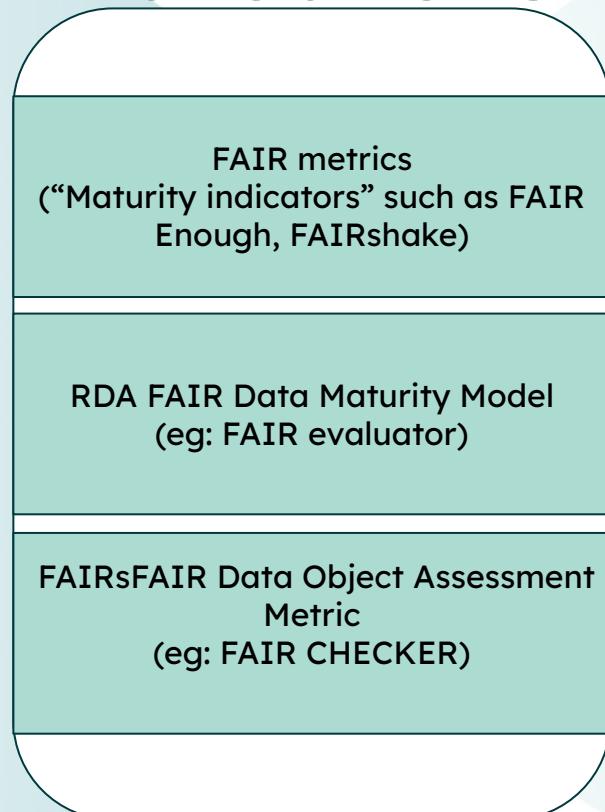
Garbage out!



Data Portal

FAIR metrics or “Indicators”

- Template for creating FAIR metrics
 - Metric ID
 - Metric Name
 - To **which** principle does the metric apply?
 - **What** is being measured?
 - **Why** should we measure it?
 - **How** do we measure it?
- RDA metric ranking
 - useful
 - important
 - essential



Data quality metrics

- Principles identify what needs to be there, but they **don't tell** what is necessary and/or sufficient
- Evaluation of FAIRness reflects the extent to which a digital resource addresses the FAIR principles
- Why FAIR metric is important?
 - Better discovery and reuse
 - Community interoperability



Scoring metric - an indicator

- Principles identify what needs to be there, but they **don't tell** what is necessary and/or sufficient
- Evaluation of FAIRness reflects the extent to which a digital resource addresses the FAIR principles
- Why FAIR metric is important?
 - Better discovery and reuse
 - Community interoperability

KPI	Measurement	F	A	I	R
Completeness	All mandatory data elements are present & captured	★		★	★
Correctness	All single data elements are correct				★
Conformity	All data elements conform to a defined standard			★	
Consistency	All data elements are consistent (meaningful record)				★
Coherence	All data elements are coherently applied across applications	★	★	★	★

Scoring metric -example

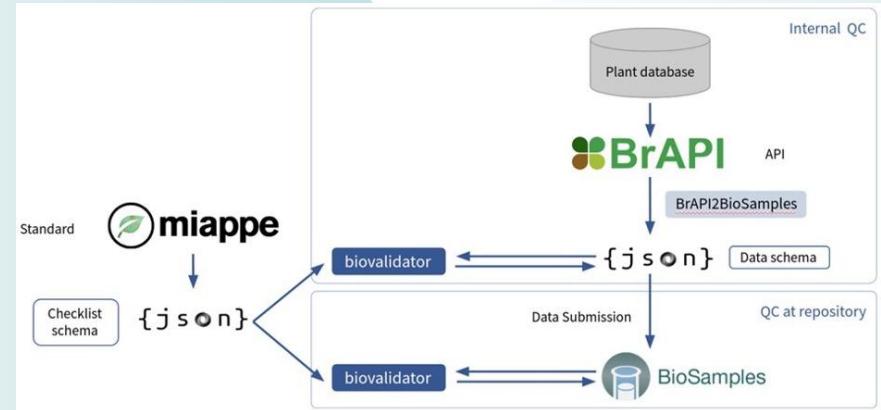
Completeness check for cell line data					
Dataset	Age	Sex	Tumor Subtype	Tumor location	Previous treatment
ABC_1	70	F	Neural	Frontal lobe	Y
ABC_2	75	-	Classical	Parietal lobe	-

Correctness check for cell line data					
Dataset	Age	Sex	Tumor Subtype	Tumor location	Previous treatment
ABC_1	80	Asian	Neural	Frontal lobe	Y
ABC_2	75.5	M	Classical	Partial lob	Y

FAIR metric assessment of cell line data		
Dataset	Completeness	Correctness
ABC_1	(5/5) 100%	(4/5) 80%
ABC_2	(3/5) 75%	(3/5) 75%

Quality Control - FAIR metric

- Data FAIRification is “need for the hour” metric for validating voluminous data (data = labor intensive, expensive)
- FAIR Validators on **semantics** and **syntactics** can favor efficient data re-use and interoperability

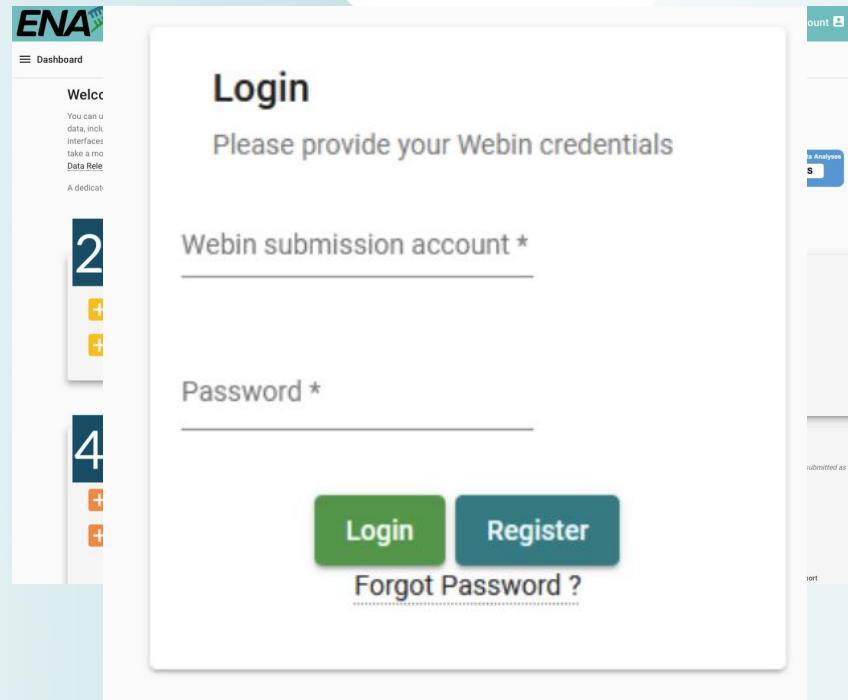


<https://academic.oup.com/bioinformatics/article/38/11/3141/6563594>

Submission to a Data Portal

-ENA

- The Webin Portal is European Nucleotide Nucleotide Archive's convenient interface for creating and reviewing submissions.
- Submission of human data through controlled access
- Functionalities
 - Submission and update of XML metadata objects
 - Spreadsheet based submission of Experiments, Samples, Runs



Metadata submission through ENA portal

- Specific requirements for metadata submissions
 - Repository contains tabular-format and xlsx spreadsheet metadata template required to submit data to

Supported metadata checklist

Real world data vs Cohort data

- RWD are important because it enables researchers to go beyond data gathered throughout a traditional controlled trials
- Can be collected from any number of cohorts or population sub-groups
- Lacks a common schema to integrate easily with external catalogue systems, difficult to harmonize datasets
- Cohort studies is that all relevant variables can be thought of in advance - clear hypothesis definition
- Data related to these variables can be accurately measured and recorded by trained study staff
- No randomization to the subgroups of interest, cause and effect relationships cannot be determined, and relationships between variables must be stated as associations

Why GHGA metadata model matters?

GHGA Metadata Model

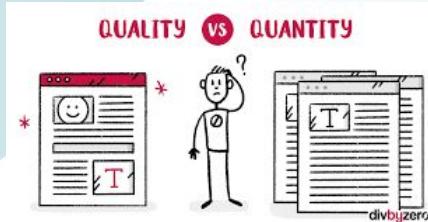
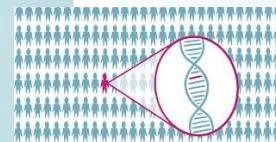
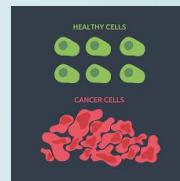
- Standardized Concepts and Standards
- Modelled using a framework that can be used to describe structure and semantics

Schema Overview



Why GHGA model matters?

- Model aligned with national and international standards
- Balanced design considering data re-use as well as data privacy
- Customizable model for different modalities
- Qualitative metadata submission with varying degrees of freedom for submitters



GHGA

Summary

- Community based initiatives play important role in establishing FAIR concepts across the scientific community
- FAIR data portals are important to increase visibility, improve reproducibility and reliability of your research
- Metrics or indicators are essential to improve the quality of data submitted to the portal - can vary according to use cases
- Balanced design of metadata models enhances the quality of data submission promoting openness and advancement of scientific research

Additional resources

- White paper: <https://doi.org/10.5281/zenodo.8341224>
- User documentation: <https://docs.ghga.de/>
- Git: <https://github.com/ghga-de/ghga-metadata-schema>