GHGA

THE GERMAN HUMAN GENOME-PHENOME ARCHIVE

GHGA Webinar

FAIR Data for FAIR Biomedical Portals

In cooperation with

nfdi

GHGA

THE GERMAN HUMAN GENOME-PHENOME ARCHIVE

Part 1: FAIR Biomedical Data

- Q&A Session

Part 2: FAIR Data Portals

- Q&A Session and Feedback Poll

In cooperation with

nfdi

FAIR Biomedical Data

GHGA

Structure

- Data and Metadata
- Biomedical Metadata

- What is FAIR…
- …and why is it important?
- How can I make my data FAIR?

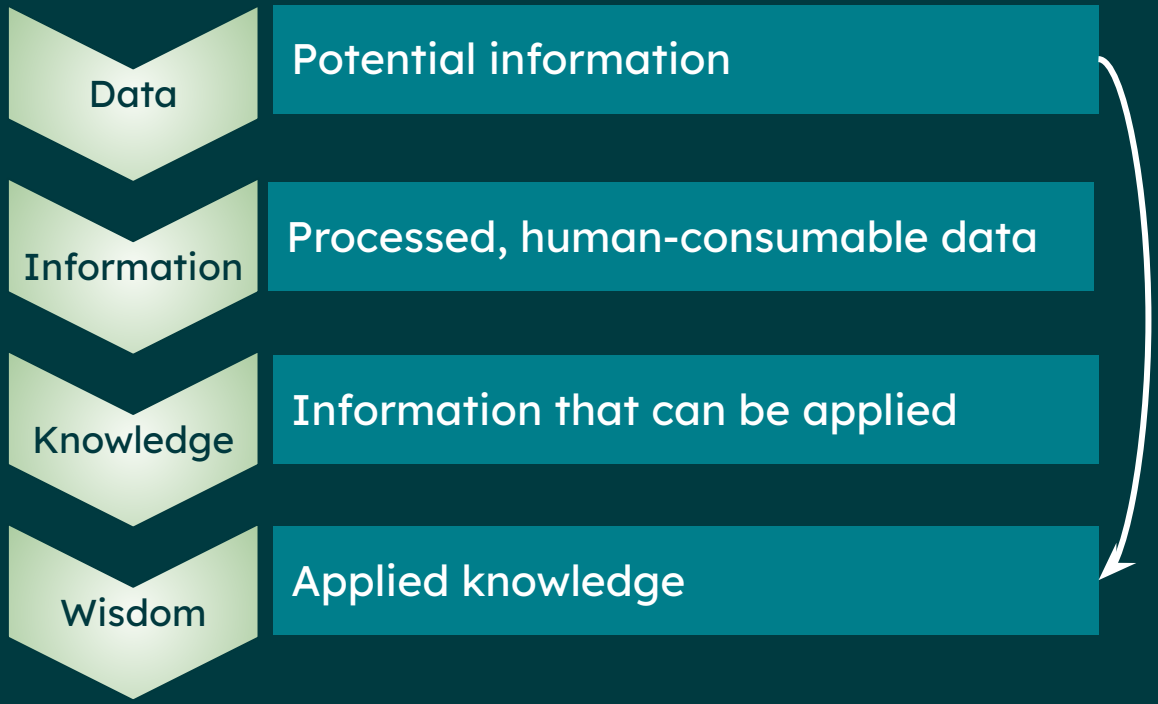
- What is a metadata model?
- Existing metadata solutions
- How to create a metadata model
- Pitfalls and things to consider

Data and Metadata

FAIR Biomedical Data

GHGA

Data and Metadata



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

GHGA

https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/

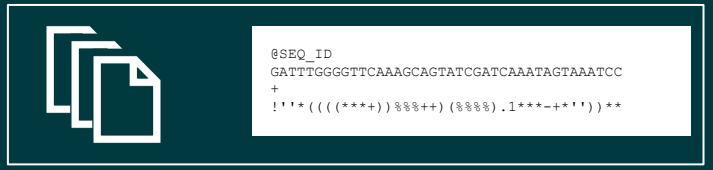
Data and Metadata

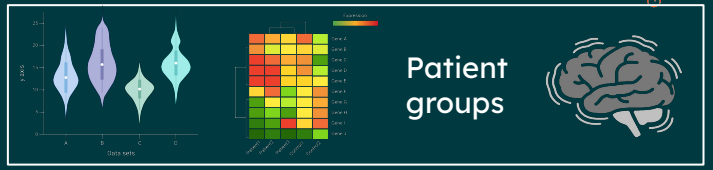
Data — Many FASTQ files

Information — Normalized and annotated gene counts

Knowledge — Differentially expressed genes

Wisdom — Possible treatments based on identified genes

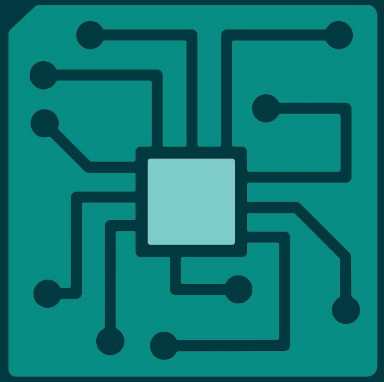




Biomedical metadata

INDIVIDUAL DATA

Demographics
Clinical information
(diagnoses, phenotype, measurements)



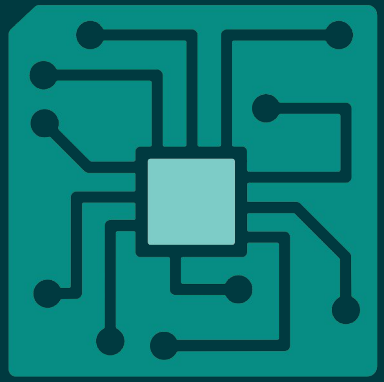
TECHNICAL DATA

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

GHGA

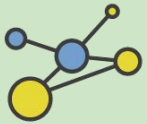
What is FAIR...

Findable

Others can find your data

Accessible

Others (know how they) can access your data



Interoperable

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

Reusable

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

GHGA

https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/

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.“

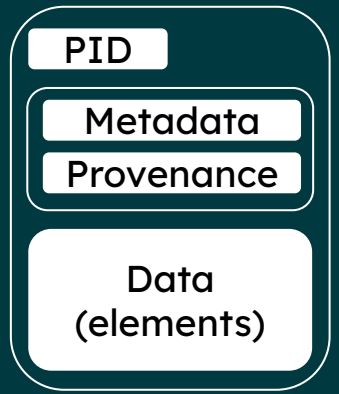
„(...) the FAIR principles put specific emphasis on enhancing the ability of **machines** to automatically find and use the data, in addition to supporting its reuse by individuals.“

Why?
- Explosion of data
- Increased international collaboration

→ Data management issues

A FAIR Digital Object

is only FAIR if all components are FAIR



https://www.nature.com/articles/sdata201618

… 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 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%)”

https://elifesciences.org/articles/67995#fig1

… 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

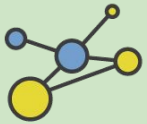
Funders and investigators must demand appropriate metadata standards to take data from foul to FAIR.

By Mark A. Musen

How can I make my data FAIR?

Findable

Accessible



Interoperable

Reusable

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

GHGA

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

GHGA

How can I make my data FAIR?

	Data Types			
Principle	**Human Bulk Omics**	**Mouse Bulk Omics**	**Mouse scRNAseq**	**Human Proteomics**
Findable	aggregated information			
Accessible	restrictions may apply			
Interoperable	existing standards to implement			
Reusable	detailed metadata might not be available			

GHGA

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

GHGA

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/

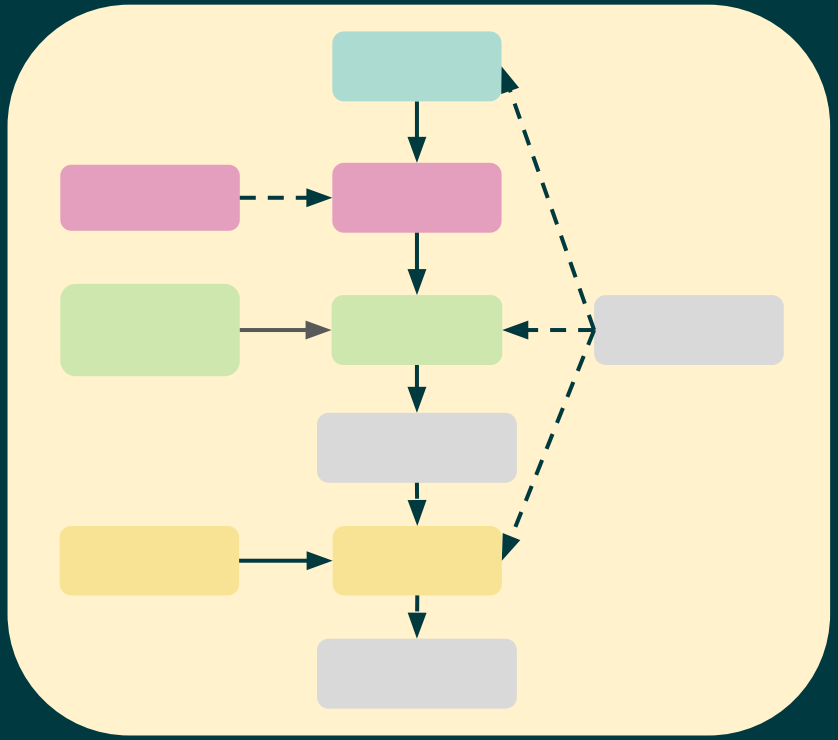
https://rdmkit.elixir-europe.org/

Metadata Models

FAIR Biomedical Data

GHGA

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 models / 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?

*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!

Ontologies and controlled vocabularies are your friends

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

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.

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.

GHGA

Additional resources

GHGA

Communities and Data Portals

GHGA

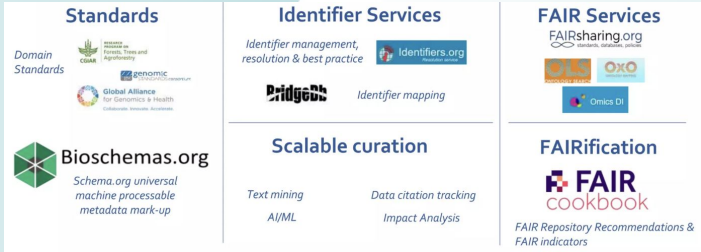
ELIXIR

- 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

Trusted domain specific repositories curated through the nodes

Supporting data repositories to be FAIR



All image source:
https://www.slideshare.net/slideshow/fair-data-bridging-from-researcher-data-management-to-elixir-archives-in-the-rdm-lifecycle/250055228#14

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

GHGA

FAIRsharing.org

- A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies

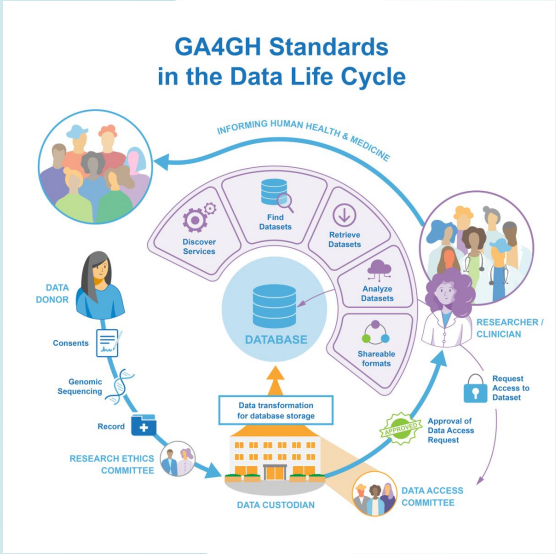
- Open standards: reporting guidelines, terminology, and exchange formats

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

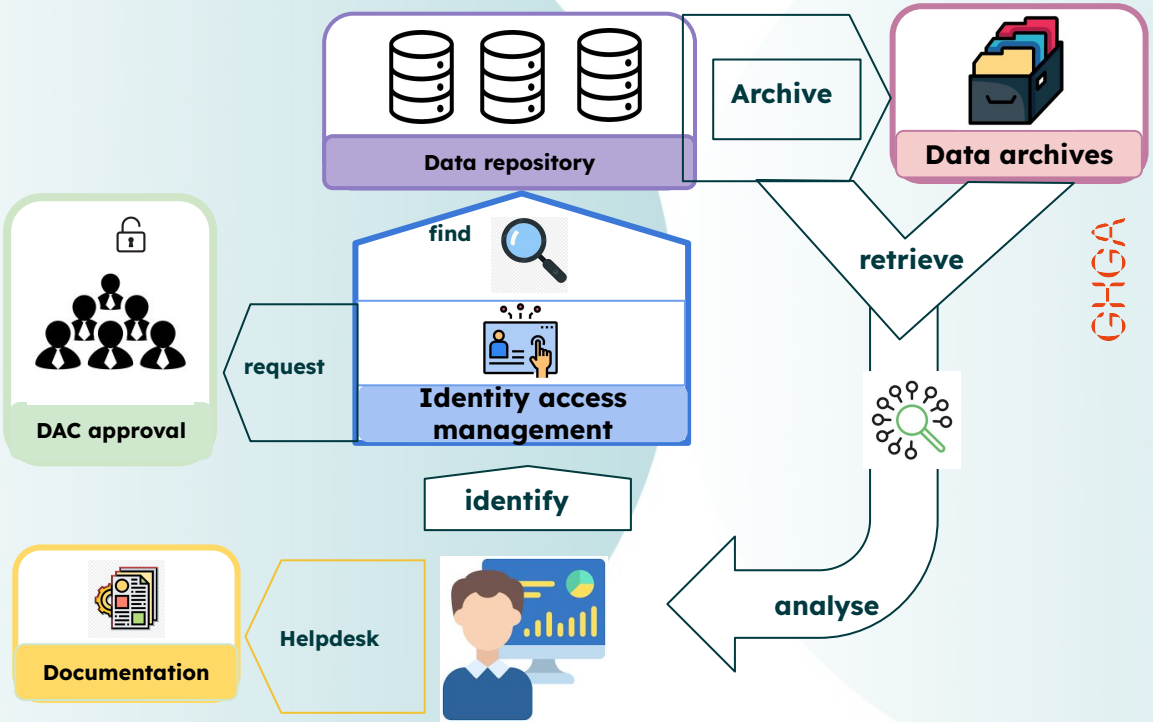


https://www.sciencedirect.com/science/article/pii/S2666979X21000367

Need for FAIR data portals

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

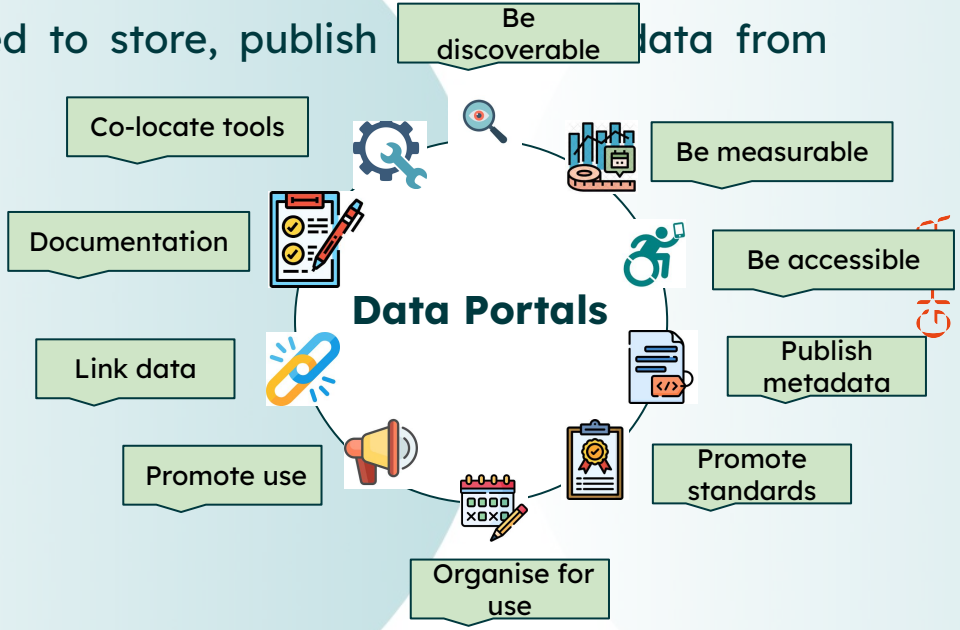
in a **reliable** and **trustworthy** way



Open Data portals

- Data portal is a platform designed to store, publish [illegible]ata from different applications and users.

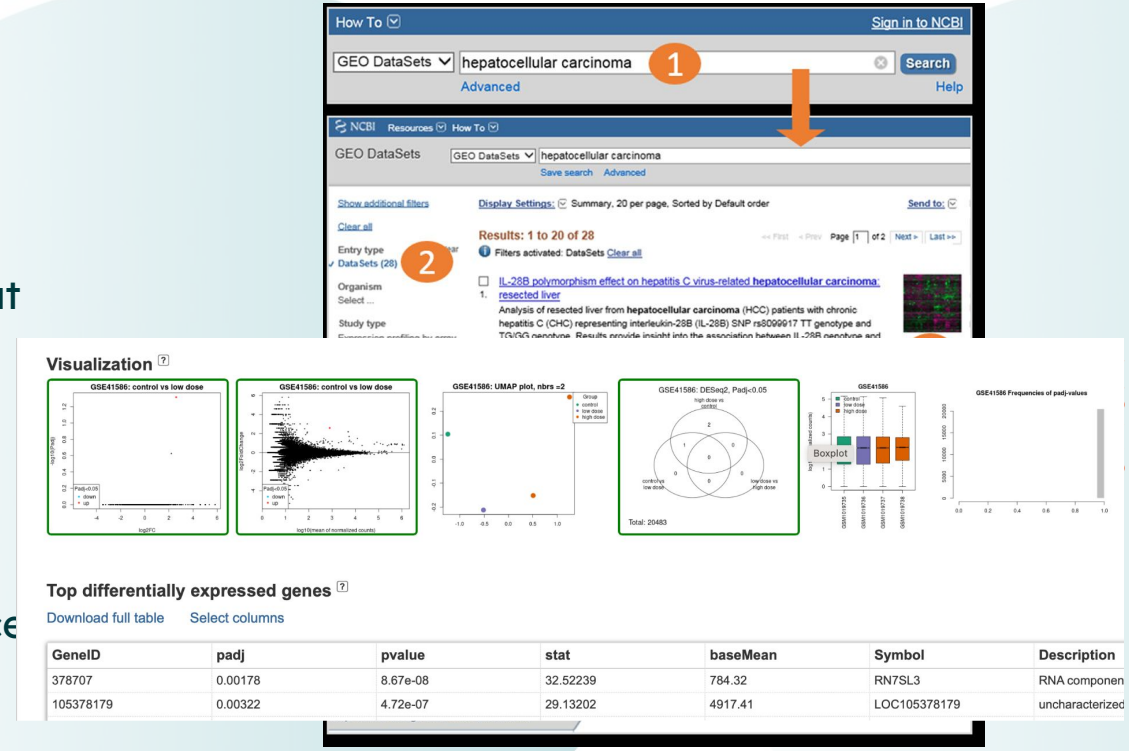
  - Easy to submit
  - Searchable
  - Transparent



NCBI GEO

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

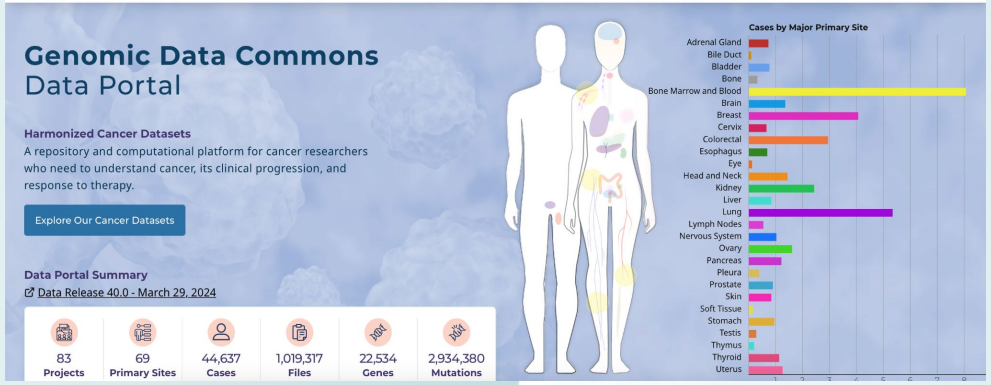
- Public archive and resource for gene expression data.

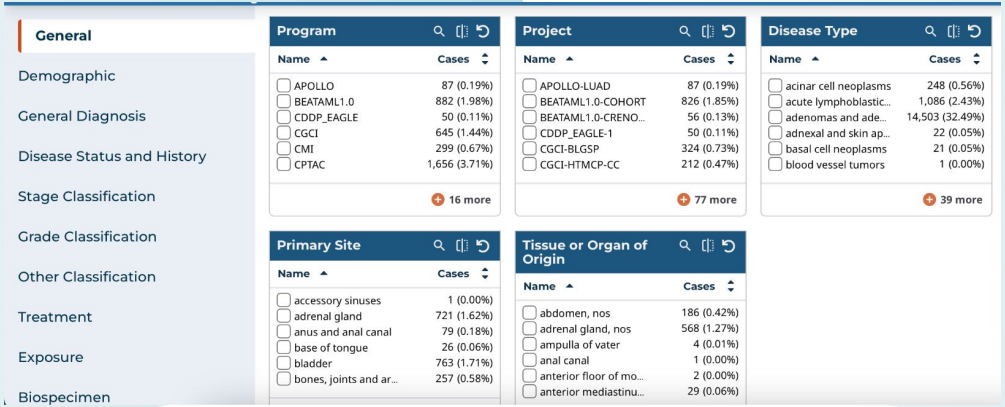


Genomic Data Commons

- GDC is a 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.

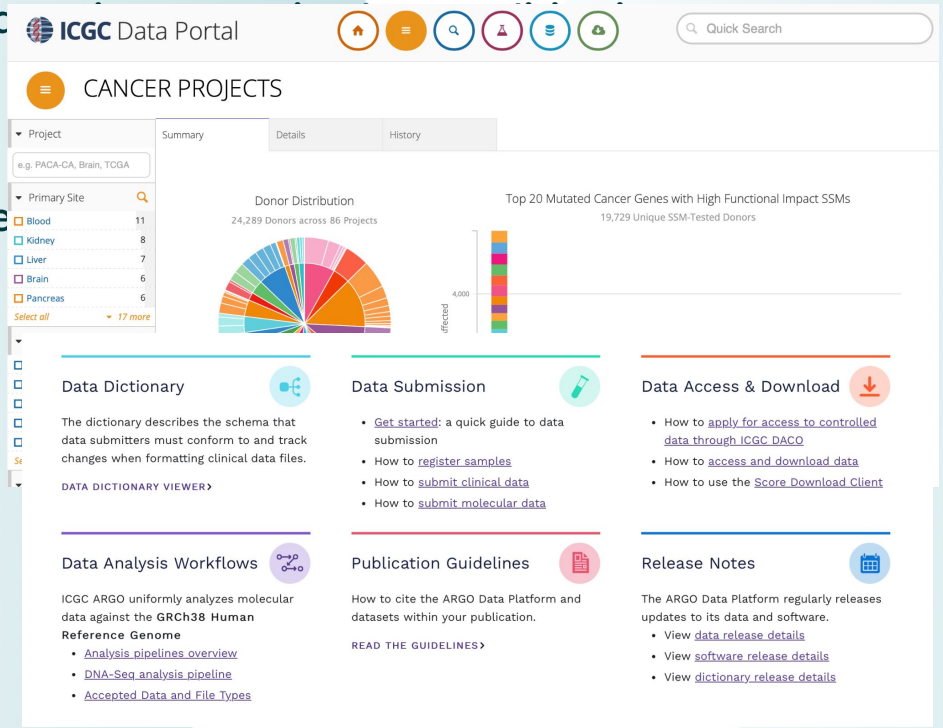




International Cancer Genome Consortium

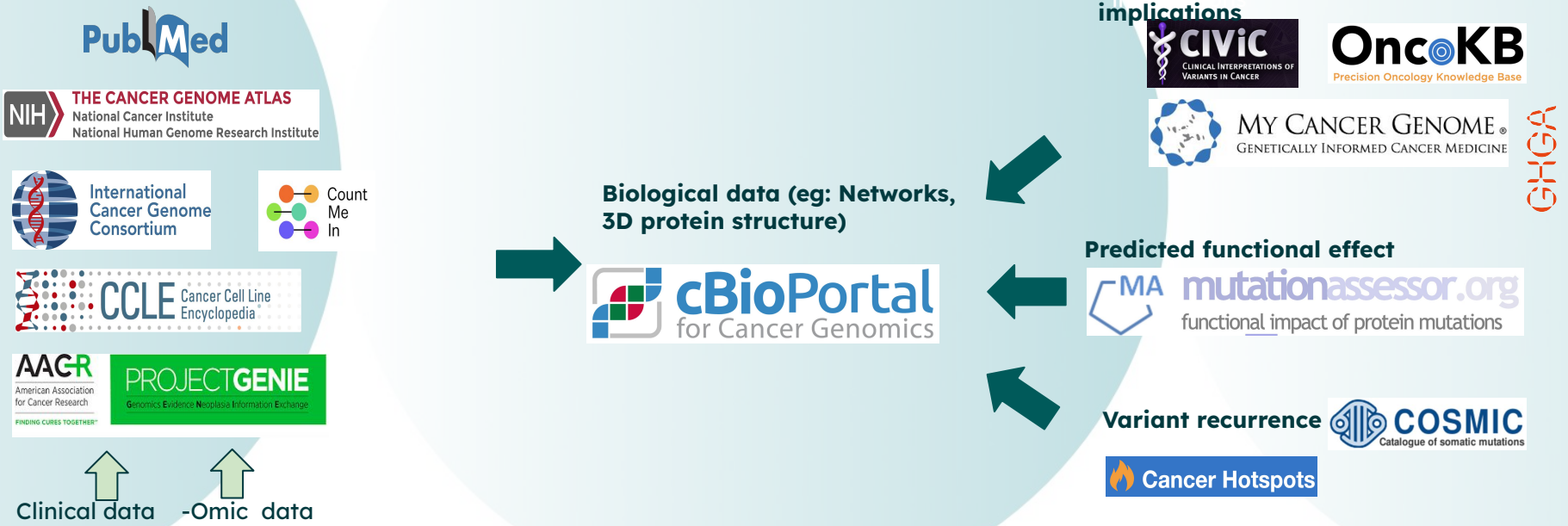
- ICGC is a collaborative effort to charc
  cancer types

- The data portal uses GUI to offer rese
  analyze data.



cBioPortal

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



Ways to assess FAIRness in data portals

GHGA

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!

Data Portal

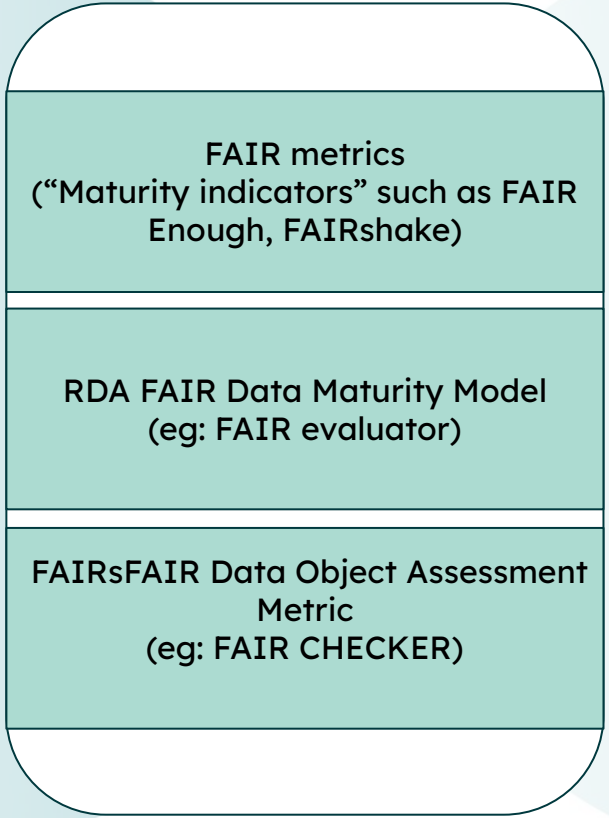
Garbage out!

GHGA

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



GHGA

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

GHGA

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	★	★	★	★

GHGA

Scoring metric -example

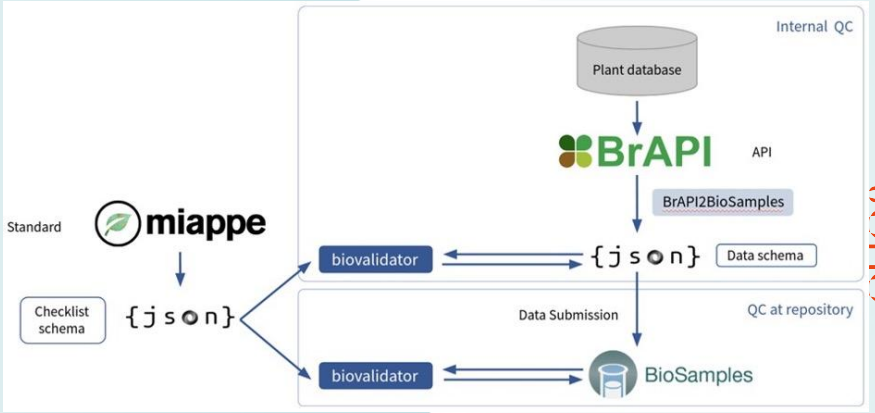
Completeness check for cell line data					
Dataset	Age	Sex	Tumor Subtype	Tumor location	Previous treatment
ABC_1	70	F	Neural	Frontal lobe	Y
ABC_2	75	-	Classical	Parietal lobe	-

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	Partal 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

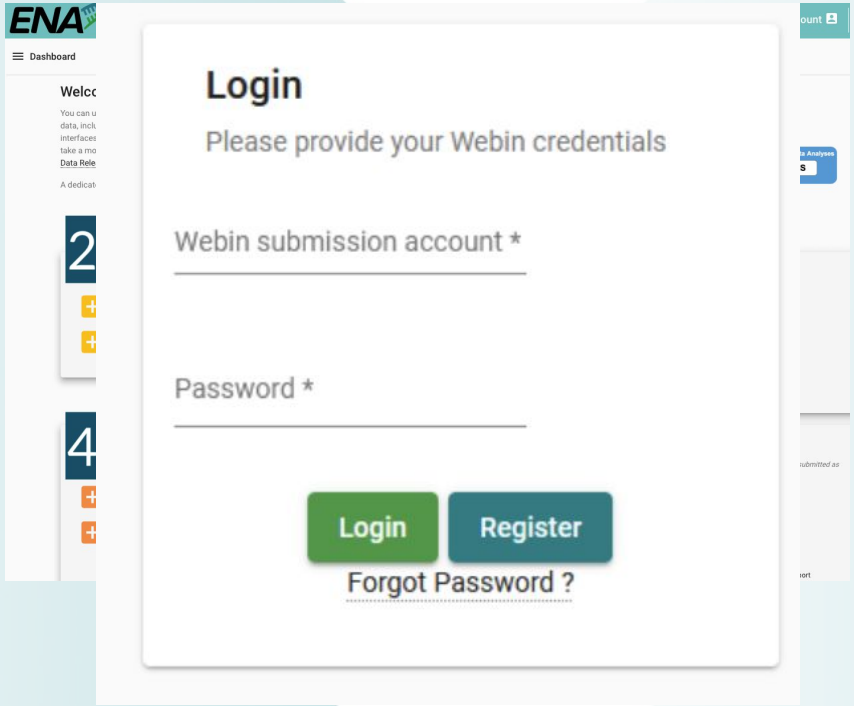
- Data FAIRification is “need for the hour” metric for validating voluminous data (data = labor intensive, expensive)
- FAIR Validators on **semantics** and **syntactics** can favor efficient data re-use and interoperability



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 templ required to submit data to EN

Supported metadata checklist

ERC000014	GSC MIxS human associated **Sample**	Genomic Standards Consortium package extension for reporting of measurements and observations obtained from the environment where the sample was obtained. By choosing the environmental package, a selection of fields can be made from a relevant subsets of the GSC terms.

A	B	C	D	E	F	G	H	I	J
alias	**title**	**taxon_id**	**sample_description**	**cell_type**	**dev_stage**	**germline**	**tissue_lib**	**tissue_type**	**isolation_source**
(Mandatory) Unique identificator for each run.	(Mandatory) Short text that can be used to call out sample records in search results or in displays.	(Mandatory) Ncbi taxonomy identifier. this is appropriate for individual organisms and some environmental samples.	(Optional) Free-form text describing the sample, its origin, and its method of isolation.	(Optional) Cell type from which the sample was obtained	(Optional) If the sample was obtained from an organism in a specific developmental stage, it is specified with this qualifier	(Optional) The sample described presented in the entry has not undergone somatic genomic rearrangement as part of an adaptive immune response; it is the	(Optional) Tissue library from which sample was obtained	(Optional) Tissue type from which the sample was obtained	(Optional) Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived

each experiment. this is used to link runs to experiments.	experiment records in searches or in displays. this element is technically optional but should be used for	study metadata)		including library was constructed.		for this library.	source material that is being sequenced.	the sequence library preparation	

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

GHGA

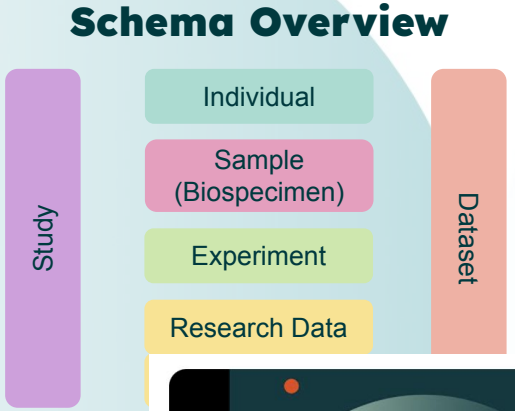
Why GHGA metadata model matters?

GHGA

GHGA Metadata Model

- Standardized Concepts and Standards

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



GHGA Webinar - Introduction to the GHGA Metadata Model w/ Karoline Mauer

Why GHGA model matters?

- Model aligned with national and international standards

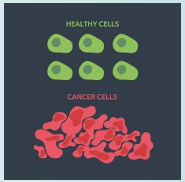
- Balanced design considering data re-use as well as data privacy

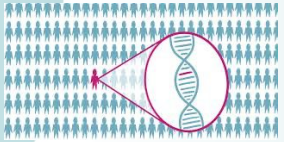
- Customizable model for different modalities

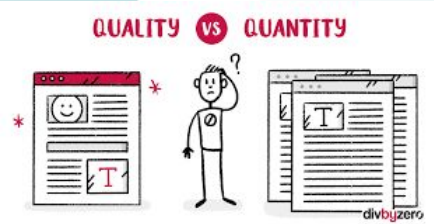
- Qualitative metadata submission with varying degrees of freedom for submitters

EUROPEAN GENOME-PHENOME ARCHIVE

Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.







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

GHGA