FAIR Principles for AI/ML Data Sharing

Yi Luo, Ph.D.

Jamie K. Teer, Ph.D.

5/23/2022

MOFFITT CANCER CENTER

Outline

Q1. **What** are FAIR Principles?

Q2. **Why** FAIR Principles are important for AI/ML data sharing?

Q3. **How** to implement FAIR Principles for AI/ML data sharing in **precision oncology**?

a. Initiatives to support interoperability and reusability in **data collection**
   i. Clinical data
   ii. Genomic data
   iii. Imaging data

b. Initiatives to support findability and accessibility in **data sharing**
   i. Network architecture
   ii. Access control

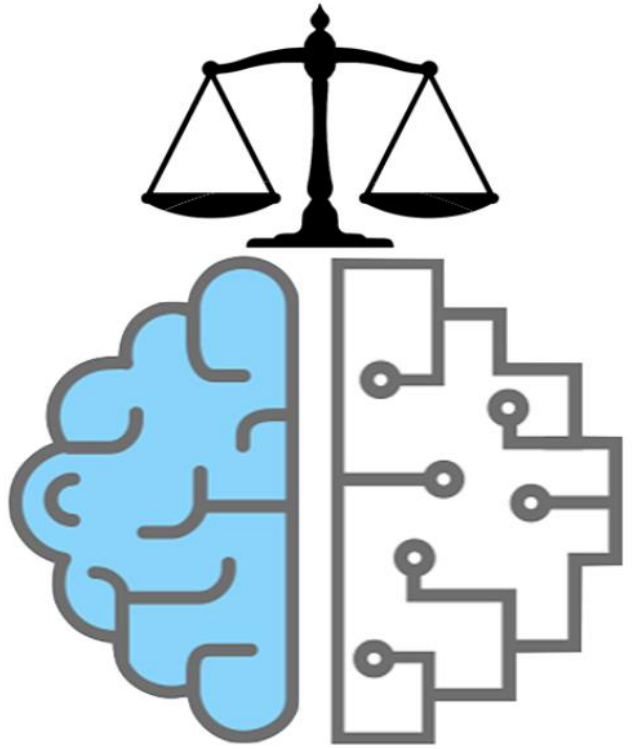
Take home messages

Question One:
What are FAIR Principles?

MOFFITT CANCER CENTER

Findability, Accessibility, Interoperability, and Reusability (FAIR)

- '**Findability**' implies data can be **found** online, typically through indexing in search engines.
- '**Accessibility**' means data can be **retrieved** directly or via an approval process.
- '**Interoperability**' imposes data to follow **standards**.
- '**Reusability**' requires the context of the data generation (metadata) is **documented** so it can be compared to or integrated with other data sets.



Charles Vesteghem, Briefings in Bioinformatics, 2020,

Findability Principle_1

F1. (Meta)data are assigned a globally unique and persistent identifier.

- **Identifiers** consist of an internet link (e.g., a uniform resource identifier (URI) that resolves to a web page).

https://www.go-fair.org/fair-principles/

Findability Principle_1

- **Metadata** is data that provides information about other data, but not the content of the data.

emlployee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Cornelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Data

Metadata

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
department_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee sti

Findability Principle_2

F1. (Meta)data are assigned a globally unique and persistent identifier.

F2. Data are described with rich metadata.

- **Metadata** can (and should) be generous and extensive, including descriptive information about the context, quality and condition, or characteristics of the data.
- **Rich metadata** allow a computer to automatically accomplish routine and tedious sorting and prioritizing tasks.

Findability Principle_3

F1. (Meta)data are assigned a globally unique and persistent identifier.

F2. Data are described with rich metadata.

F3. Metadata clearly and explicitly include the identifier of the data they describe.

- The metadata and the dataset they describe are usually **separate files**. The association between them should be made explicit by mentioning a dataset's globally unique and persistent identifier in the metadata.
- Many **repositories** will generate globally unique and persistent identifiers for deposited datasets that can be used for this purpose.

Findability Principle_4

F1. (Meta)data are assigned a globally unique and persistent identifier.

F2. Data are described with rich metadata (defined by R1 below).

F3. Metadata clearly and explicitly include the identifier of the data they describe.

F4. (Meta)data are registered or indexed in a searchable resource.

- If the **availability** of a digital resource such as a dataset, service or repository is not known, then nobody (and no machine) can discover it. Identifiers and rich metadata descriptions alone will not ensure '**findability**' on the internet.
- There are many ways in which digital resources can be made discoverable, including **indexing**.

Accessibility Principle_1

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol.

- Most users of the internet retrieve data by ‘clicking on a link’. This is a high-level interface to a low-level protocol called **Transmission Control Protocol** (TCP), that the computer executes to load data in the user’s web browser.

- FAIR data retrieval should be mediated **without** specialized tools or communication methods. This principle focuses on how data and metadata can be retrieved from their identifiers.

Accessibility Principle_1.1

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 The protocol is open, free, and universally implementable.

- To **maximize data reuse**, the protocol should be free (no-cost) and open (-sourced) and thus globally implementable to facilitate data retrieval.
- Anyone with a computer and an internet connection can access at least the metadata. Hence, this criterion will impact your choice of the **repository** where you will share your data.

Accessibility Principle_1.2

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 The protocol is open, free, and universally implementable.

A1.2 The protocol allows for an authentication and authorization procedure, where necessary.

- The 'A' in FAIR does not necessarily mean '**open**' or '**free**'. Rather, it implies that one should provide the exact conditions under which the data are accessible. Hence, even **heavily protected and private data** can be FAIR.
- Ideally, accessibility is specified in such a way that a **machine** can **automatically** understand the requirements, and then either **automatically** execute the requirements or alert the user to the requirements.

Accessibility Principle_2

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 The protocol is open, free, and universally implementable.

A1.2 The protocol allows for an authentication and authorization procedure, where necessary.

A2. Metadata are accessible, even when the data are no longer available.

- Datasets tend to degrade or disappear over time because there is a **cost** to maintaining an online presence for data resources.
- Storing the metadata generally is much **easier** and **cheaper**. Hence, metadata should persist even when the data are no longer sustained.

Interoperability Principle_1

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- Humans should be able to exchange and interpret each other's data. Also, data should be **readable for machines** without the need for specialized or ad hoc algorithms, translators, or mappings.
- Each computer system at least has knowledge of the other system's **data exchange formats**. It is critical to use
  (1) commonly used controlled **vocabularies**, **ontologies**
  (2) a well-defined **framework** to describe and structure (meta)data.

Interoperability Principle_2

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles.

- The controlled vocabulary used to describe datasets needs to be documented and resolvable using **globally unique and persistent identifiers**. This documentation needs to be easily findable and accessible by anyone who uses the dataset.

Interoperability Principle_3

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles.

I3. (Meta)data include qualified references to other (meta)data.

- A qualified reference is a **cross-reference** that explains its intent.
- In particular, the scientific links between the datasets need to be described. Furthermore, all datasets need to be properly **cited**.

Reusability Principle_1

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.

- The data publisher should provide not just metadata that allows **discovery**, but also metadata that richly describes the **context** under which the data was generated.
- This may include the experimental protocols, the manufacturer and brand of the machine or sensor that created the data, etc.

Reusability Principle_1.1

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.

R1.1. (Meta)data are released with a clear and accessible data usage license.

- 'I' principles covered elements of **technical interoperability**. R1.1 is about **legal interoperability**. What usage rights do you attach to your data? This should be described clearly.
- Clarity of **licensing status** will become more important with automated searches involving more licensing considerations.

Reusability Principle_1.2

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.

R1.1. (Meta)data are released with a clear and accessible data usage license.

R1.2. (Meta)data are associated with detailed provenance.

- Include a description of the **workflow** that led to your data: Who generated or collected it? How has it been processed? Has it been published before? Ideally, this workflow is described in a machine-readable format.

Reusability Principle_1.3

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.

R1.1. (Meta)data are released with a clear and accessible data usage license.

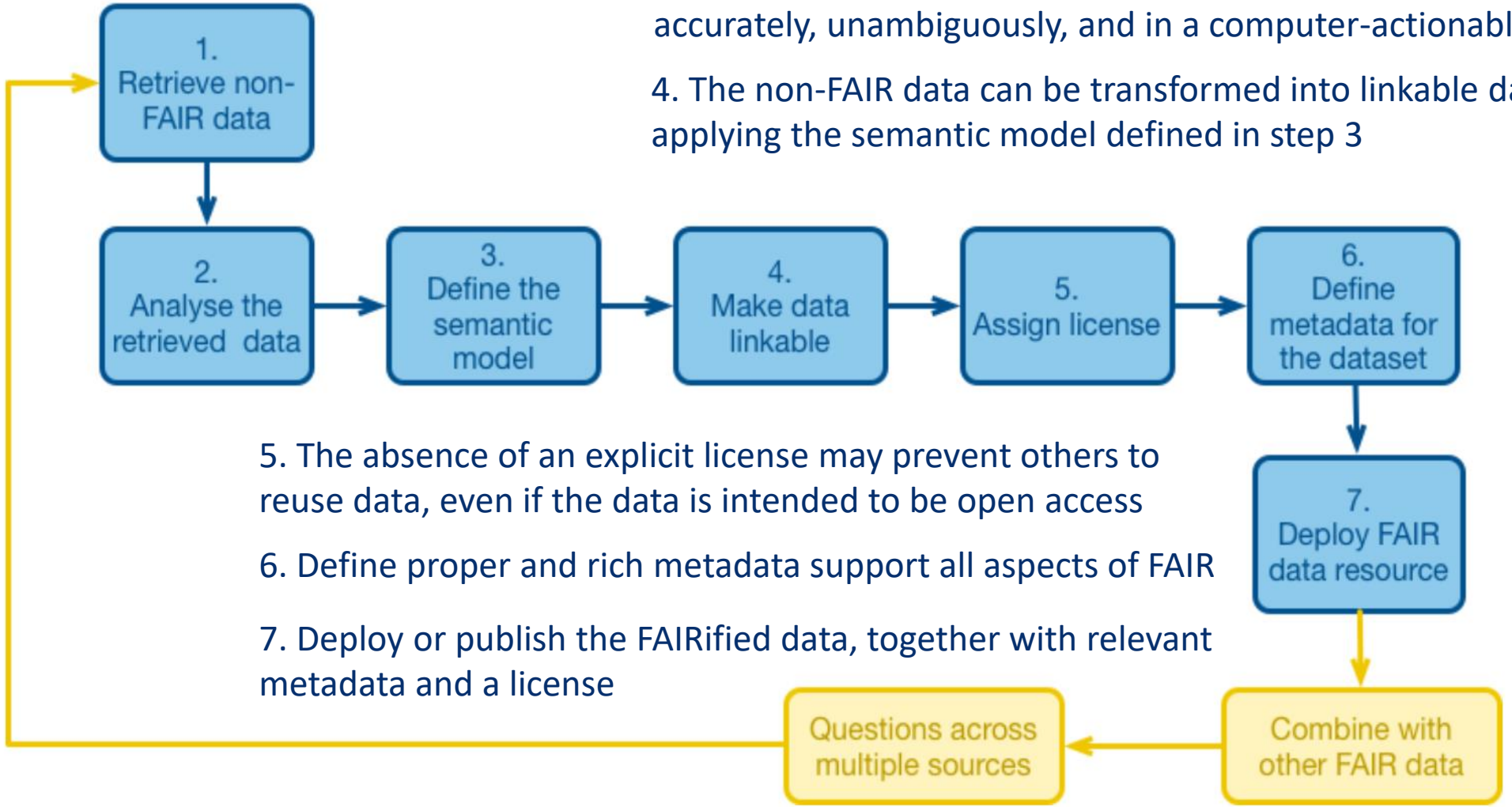
R1.2. (Meta)data are associated with detailed provenance.

R1.3. (Meta)data meet domain-relevant community standards.

- It is easier to reuse data sets if they are **similar**: same type of data, data organized in a standardized way, etc.
- Many communities have minimal **information standards**. FAIR data should at least meet those standards.

FAIRification Process

1. Gain access to the data to be FAIRified
2. Inspect the content of the data
3. Describes the meaning of entities and relations in the dataset accurately, unambiguously, and in a computer-actionable way
4. The non-FAIR data can be transformed into linkable data by applying the semantic model defined in step 3



5. The absence of an explicit license may prevent others to reuse data, even if the data is intended to be open access
6. Define proper and rich metadata support all aspects of FAIR
7. Deploy or publish the FAIRified data, together with relevant metadata and a license

https://www.go-fair.org/fair-principles/

Question Two:
Why FAIR Principles are Important for Data Sharing?

MOFFITT CANCER CENTER

Benefits for Implementation of FAIR Principles

The entire process of running data through the **value chain** from acquisition, semantic alignment, integration to analytics to generate insights will become streamlined and, as a result, more effective.

- **Accelerating** innovation owing to availability of FAIR data for primary use and secondary reuse;
- **Developing** more-segmented or –personalized medicines by exploiting FAIR real-word data to match best treatment to relevant patient cohorts;
- **Enabling** data sharing and collaborations across institutions and companies.

John Wise, Drug Discovery Today, 2019

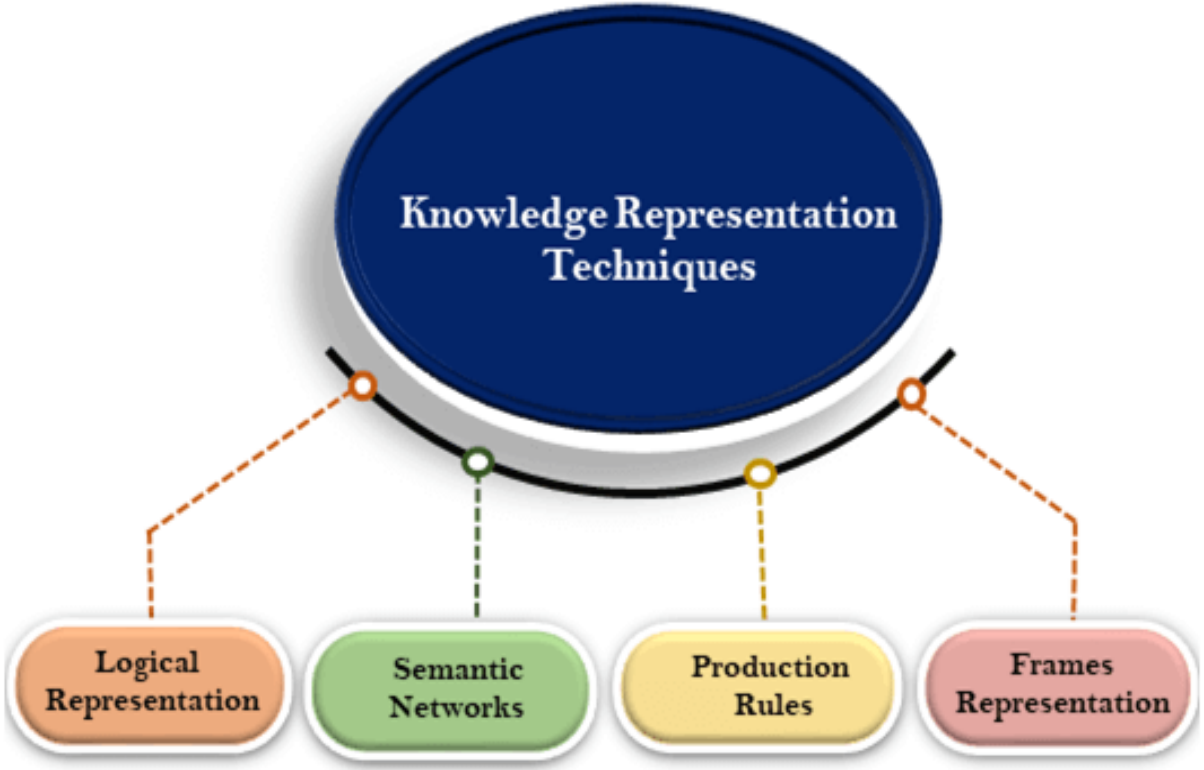
Data Governance

- **Data governance** is a collection of processes, roles, policies, standards, and metrics that ensure the **effective** and **efficient** use of information in enabling an organization to achieve its goals.

- The FAIR principles provide an important but **not exhaustive** foundation to define it.

Knowledge Representation

- **Knowledge representation** is the ability for computer systems to understand information about the world with sufficient accuracy to utilize that information for an intelligent purpose.
- Successful deployment of FAIR will require a **standardized information architecture**, which helps reach a robust consensus on the ontologies in capturing specific types of data.

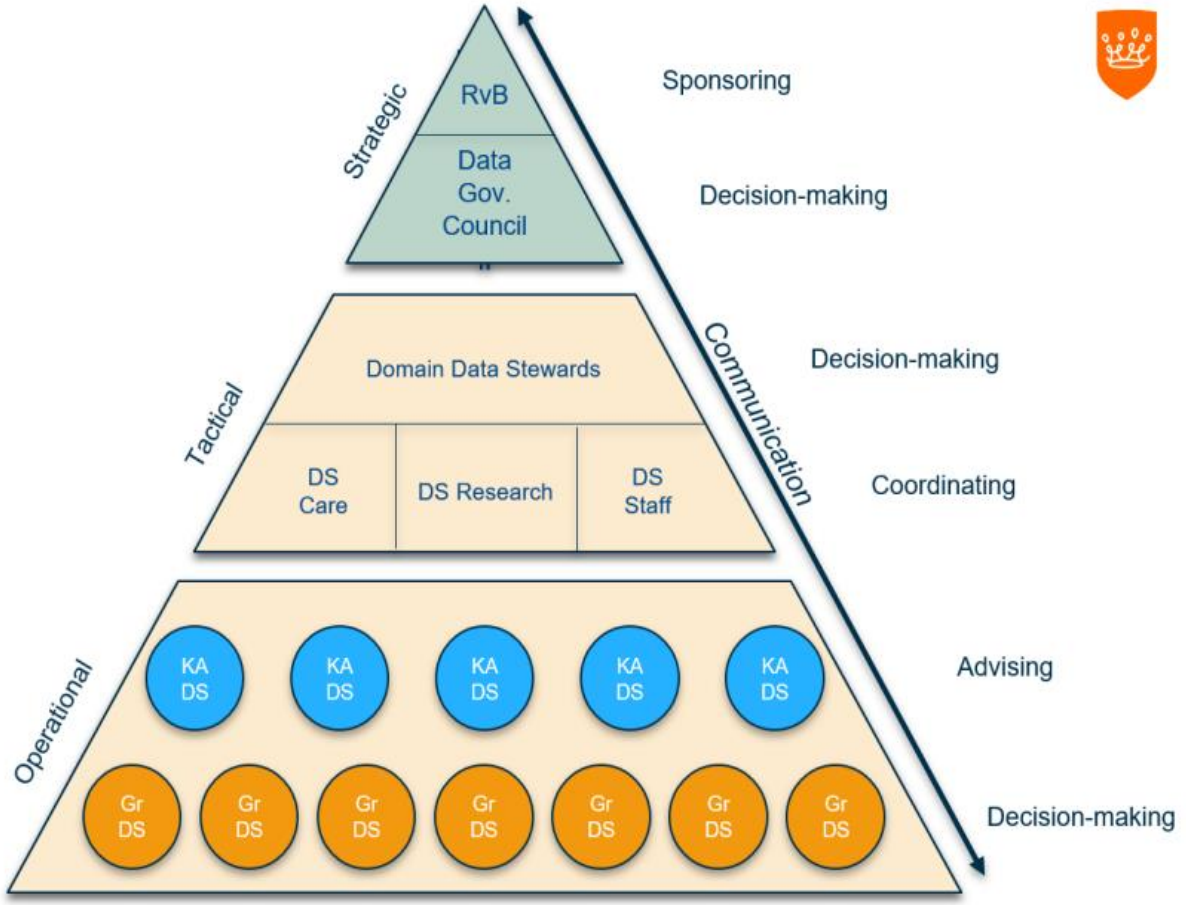


Javatpoint.com, Techniques of knowledge representation

Data Stewardship

- **Data Stewardship** has the role of ensuring an adequate level of data and metadata quality throughout the organization utilizing data governance processes.

- Data integration challenges and quality issues usually emerge late at the time of analysis and reuse; the best point of action for quality control is at the point of **data collection**.



FAIR Data Principles in Precision Oncology

- The goal of precision medicine is to take a detailed view of each patient and their cancer to tailor their treatment accordingly. Research has recently shown that cancer is so **heterogeneous** that single research centers cannot produce enough data to fit prognostic and predictive models of sufficient accuracy. **Data sharing** in precision oncology is therefore of utmost importance.

- However, in addition to **privacy** and **ethical** issues, various local and national health care systems and reporting traditions are often **incompatible**, making it **complicated**, **expensive**, and **time-consuming** to aggregate data from different sources due to the amount of data management involved. Then various **initiatives** have been launched to tackle these issues

Charles Vesteghem, et al, Briefings in Bioinformatics, 2020

FAIR Principles: the Moffitt Bioinformatics Experience

- **Findable**
  - Sample identifiers should be unique, machine sortable, non-PHI
  - Should indicate project, PI, sample number. (not just PI_1, PI_2, …)
  - Organization of analysis metadata and processes (including tools and versions used) helps answer later questions
  - At Moffitt, patient samples should always be linkable to a patient!
    - Can be done in a de-identified manner
    - Allows association of molecular and other data with clinical information
- **Accessible**
  - Metadata should be in a standard format to allow easy searching
  - Locally stored data must be organized in a way that is **findable** in the future!

MOFFITT CANCER CENTER

FAIR Principles: the Moffitt Bioinformatics Experience

- **Interoperable**
  - Used standardized terms, ontology, data formats, references to enable analysis with a greater number of tools.
  - Standardization important for data reuse, but also for impactful application development.
  - Always include reference/acknowledgement of others' data!!!
- **Reusable**
  - When depositing data, include all fields needed to reproduce work.
  - When submitting patient data, protect our patients by using protected data repositories. Data are still available, but protected from misuse.
  - The exact analysis used is important to understand and interpret the data. Stay tune for Best Practices lecture!

MOFFITT CANCER CENTER

Question Three:

How to Implement FAIR Principles for AI/ML Data Sharing in **Precision Oncology**?

a. Initiatives to support interoperability and reusability in data collection

i. Clinical Data
ii. Genomic Data
iii. Imaging Data

MOFFITT CANCER CENTER

Clinical Data

Clinical data encompass the information about **patient status** and **disease phenotype**.

- The **patient status** includes demographic information, medication, comorbidities, exposures, blood test results, and treatment information.
- The **disease phenotype** is characterized by morphology and topography.
  - Morphology details the cellular structure of the cancer.
  - Topography defines the morphology's location.
- Usually, these data are collected by healthcare personnel and stored in **electronic health records** (EHRs) or in a research and clinical trials context, such as **case report forms** (CRFs).

MOFFITT CANCER CENTER

Charles Vesteghem, et al, Briefings in Bioinformatics, 2020

Major Initiatives for Clinical Data

- **Data Structure Models**
  - EHR, FHIR, CRF, REDCap, GDC
- **Ontologies**
  - caDSR, LOINC, SNOMED CT
- **Classifications**
  - ICDs, ICD-O, ATC
- **Provenance Information**
  - FAIR-Health

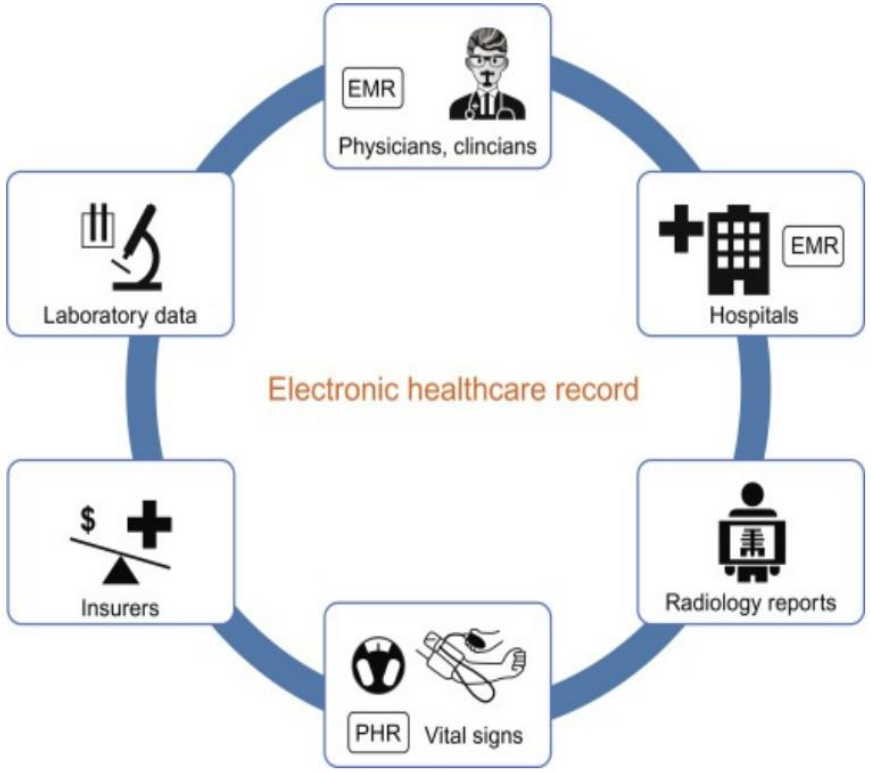
MOFFITT CANCER CENTER

Charles Vesteghem, et al, Briefings in Bioinformatics, 2020

Data Structure Models - EHR

Electronic Health Record (EHR)

- An EHR is an electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the **key administrative clinical data** relevant to that person's care under a particular provider, such as demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports.



S. Rakesh Kumar, 2019

- Several countries have largely solved the problem at a regional level but is still facing **interoperability** issues at the national level. Large efforts are needed to converge to a more broadly accepted open standard.

Data Structure Models - FHIR

Fast Healthcare Interoperability Resources (FHIR)

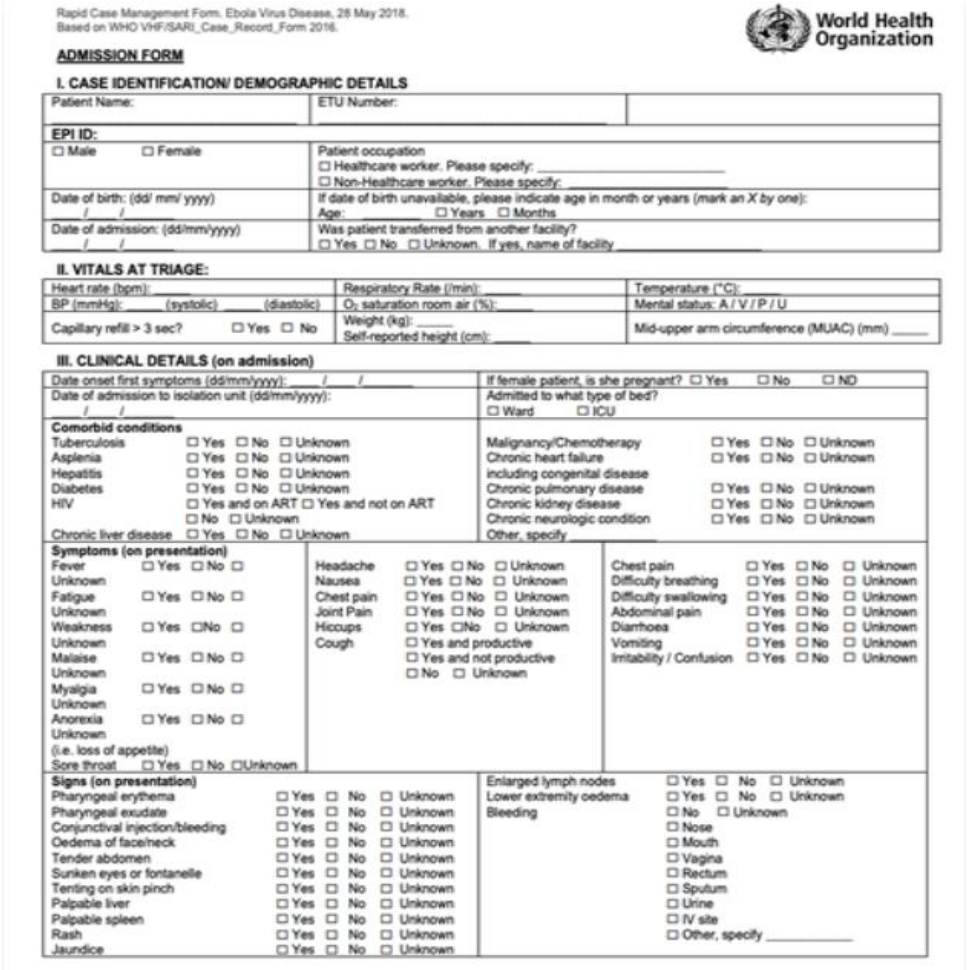
- The FHIR is a standard describing data formats and elements (known as "resources") and an application programming interface (API) for exchanging **electronic health records** (EHR). This standard is related to another older initiative from Health Level Seven, focusing exclusively on clinical data.
- Due to its **specificity**, this format might be cumbersome outside of an **EHR** context. Furthermore, the format supports a multitude of data types but does not provide guidance on what to share.

Victor Savevski, 2019

Data Structure Models - CRF

Case Report Form (CRF)

- A CRF is a paper or **electronic questionnaire** specifically used in clinical trial research. It is the tool used by the sponsor of the clinical trial to collect data from each participating patient. The sponsor of the clinical trial develops the CRF to collect the specific data they need in order to test their **hypotheses** or answer their **research questions**.



Rapid Case Management Form. Ebola Virus Disease. 28 May 2018.
Based on WHO VHF/SARI_Case_Record_Form 2016.

World Health Organization

ADMISSION FORM

I. CASE IDENTIFICATION/ DEMOGRAPHIC DETAILS

Patient Name: | ETU Number:

EPI ID:

☐ Male ☐ Female

Patient occupation
☐ Healthcare worker. Please specify: _____
☐ Non-Healthcare worker. Please specify: _____

Date of birth: (dd/ mm/ yyyy) ___/___/___

If date of birth unavailable, please indicate age in month or years (*mark an X by one*):
Age: _____ ☐ Years ☐ Months

Date of admission: (dd/mm/yyyy) ___/___/___

Was patient transferred from another facility?
☐ Yes ☐ No ☐ Unknown. If yes, name of facility _____

II. VITALS AT TRIAGE:

Heart rate (bpm): _____ | Respiratory Rate (/min): _____ | Temperature (°C): _____

BP (mmHg): _____ (systolic) _____ (diastolic) | O_2 saturation room air (%): _____ | Mental status: A / V / P / U

Capillary refill > 3 sec? ☐ Yes ☐ No | Weight (kg): _____ Self-reported height (cm): _____ | Mid-upper arm circumference (MUAC) (mm) _____

III. CLINICAL DETAILS (on admission)

Date onset first symptoms (dd/mm/yyyy): ___/___/___ | If female patient, is she pregnant? ☐ Yes ☐ No ☐ ND

Date of admission to isolation unit (dd/mm/yyyy): ___/___/___ | Admitted to what type of bed? ☐ Ward ☐ ICU

Comorbid conditions

Tuberculosis ☐ Yes ☐ No ☐ Unknown
Asplenia ☐ Yes ☐ No ☐ Unknown
Hepatitis ☐ Yes ☐ No ☐ Unknown
Diabetes ☐ Yes ☐ No ☐ Unknown
HIV ☐ Yes and on ART ☐ Yes and not on ART ☐ No ☐ Unknown
Chronic liver disease ☐ Yes ☐ No ☐ Unknown

Malignancy/Chemotherapy ☐ Yes ☐ No ☐ Unknown
Chronic heart failure including congenital disease ☐ Yes ☐ No ☐ Unknown
Chronic pulmonary disease ☐ Yes ☐ No ☐ Unknown
Chronic kidney disease ☐ Yes ☐ No ☐ Unknown
Chronic neurologic condition ☐ Yes ☐ No ☐ Unknown
Other, specify _____

Symptoms (on presentation)

Fever ☐ Yes ☐ No ☐ Unknown
Fatigue ☐ Yes ☐ No ☐ Unknown
Weakness ☐ Yes ☐No ☐ Unknown
Malaise ☐ Yes ☐ No ☐ Unknown
Myalgia ☐ Yes ☐ No ☐ Unknown
Anorexia ☐ Yes ☐ No ☐ Unknown (i.e. loss of appetite)
Sore throat ☐ Yes ☐ No ☐Unknown

Headache ☐ Yes ☐ No ☐ Unknown
Nausea ☐ Yes ☐ No ☐ Unknown
Chest pain ☐ Yes ☐ No ☐ Unknown
Joint Pain ☐ Yes ☐ No ☐ Unknown
Hiccups ☐ Yes ☐No ☐ Unknown
Cough ☐ Yes and productive ☐ Yes and not productive ☐ No ☐ Unknown

Chest pain ☐ Yes ☐ No ☐ Unknown
Difficulty breathing ☐ Yes ☐ No ☐ Unknown
Difficulty swallowing ☐ Yes ☐ No ☐ Unknown
Abdominal pain ☐ Yes ☐ No ☐ Unknown
Diarrhoea ☐ Yes ☐ No ☐ Unknown
Vomiting ☐ Yes ☐ No ☐ Unknown
Irritability / Confusion ☐ Yes ☐ No ☐ Unknown

Signs (on presentation)

Pharyngeal erythema ☐ Yes ☐ No ☐ Unknown
Pharyngeal exudate ☐ Yes ☐ No ☐ Unknown
Conjunctival injection/bleeding ☐ Yes ☐ No ☐ Unknown
Oedema of face/neck ☐ Yes ☐ No ☐ Unknown
Tender abdomen ☐ Yes ☐ No ☐ Unknown
Sunken eyes or fontanelle ☐ Yes ☐ No ☐ Unknown
Tenting on skin pinch ☐ Yes ☐ No ☐ Unknown
Palpable liver ☐ Yes ☐ No ☐ Unknown
Palpable spleen ☐ Yes ☐ No ☐ Unknown
Rash ☐ Yes ☐ No ☐ Unknown
Jaundice ☐ Yes ☐ No ☐ Unknown

Enlarged lymph nodes ☐ Yes ☐ No ☐ Unknown
Lower extremity oedema ☐ Yes ☐ No ☐ Unknown
Bleeding ☐ No ☐ Unknown
☐ Nose
☐ Mouth
☐ Vagina
☐ Rectum
☐ Sputum
☐ Urine
☐ IV site
☐ Other, specify _____

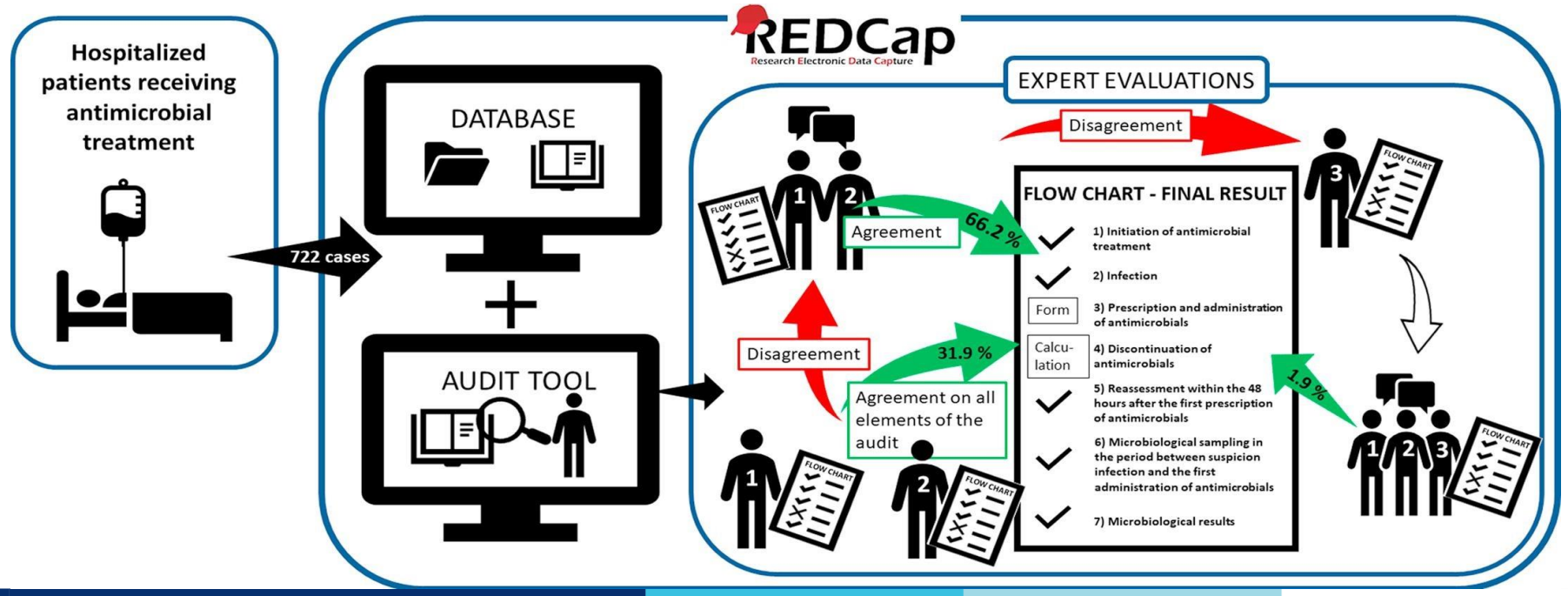
World Health Organization, 2018

- Before being sent to the sponsor, the CRFs are usually **de-identified** (not traceable to the patient) by removing the patient's name, medical record number, etc., and giving the patient a **unique study number**.

Data Structure Models - REDCap

Research Electronic Data Capture (REDCap)

- The REDCap solution allows one to design **CRFs**, but the emphasis is more on **implementation** than on good practices, as it is not designed for specifications but for actual data collection.
- The problem remains that these solutions do not provide a clear guideline on **what to collect** in the context of data sharing.



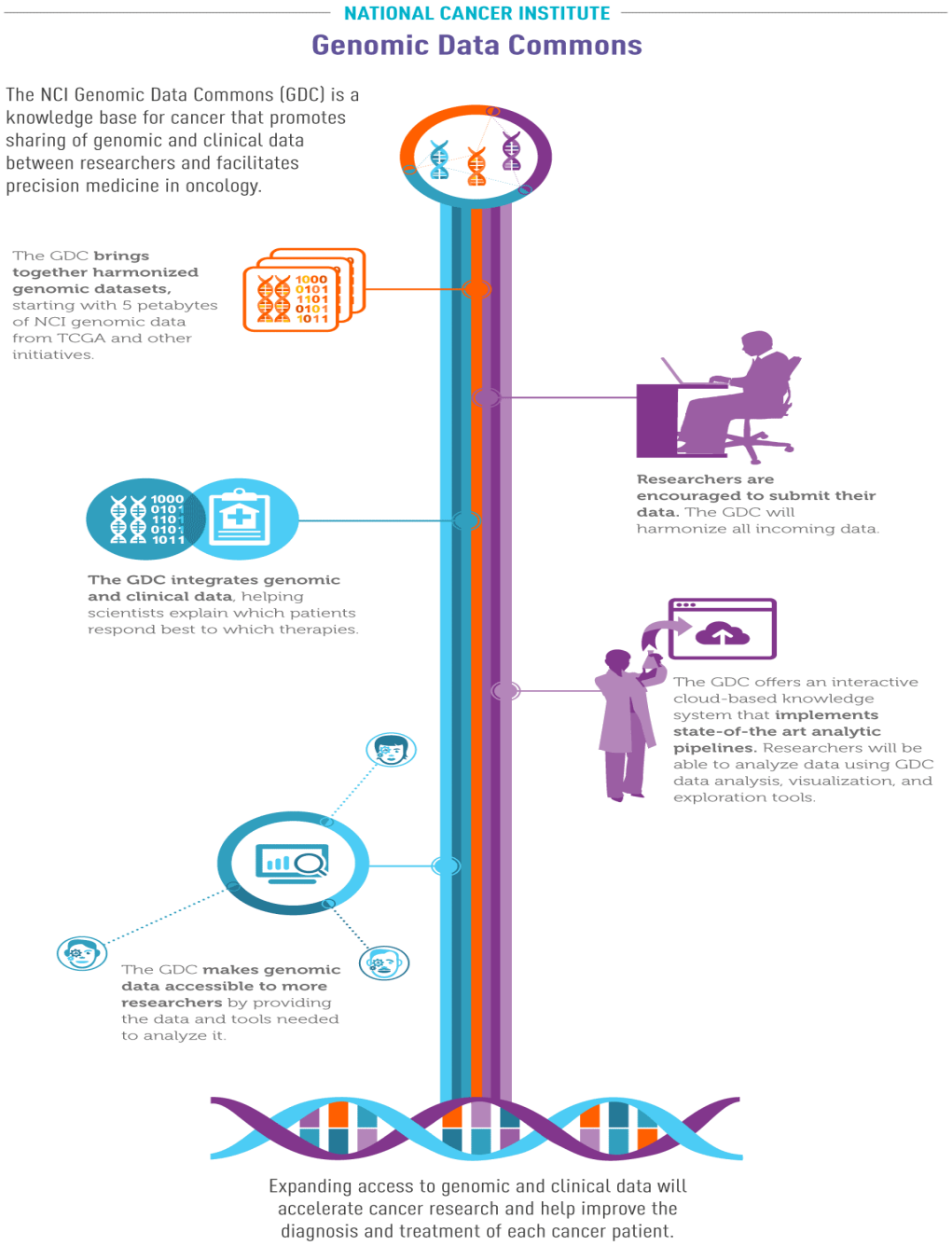
Signe H.Kragelund. 2018

Data Structure Models - GDC

Genomic Data Commons (GDC)

- The GDC's goal is to share linked clinical and genomic data from the **Therapeutically Applicable Research to Generate Effective Treatments** (TARGET) and **The Cancer Genome Atlas** (TCGA) projects. This is the largest public data repository to date linking these two types of data.

- A major **accomplishment** of the GDC was its ability to successfully gather and share data from disparate sources in a **harmonized** way as detailed harmonization requirements and procedures were designed for that purpose.

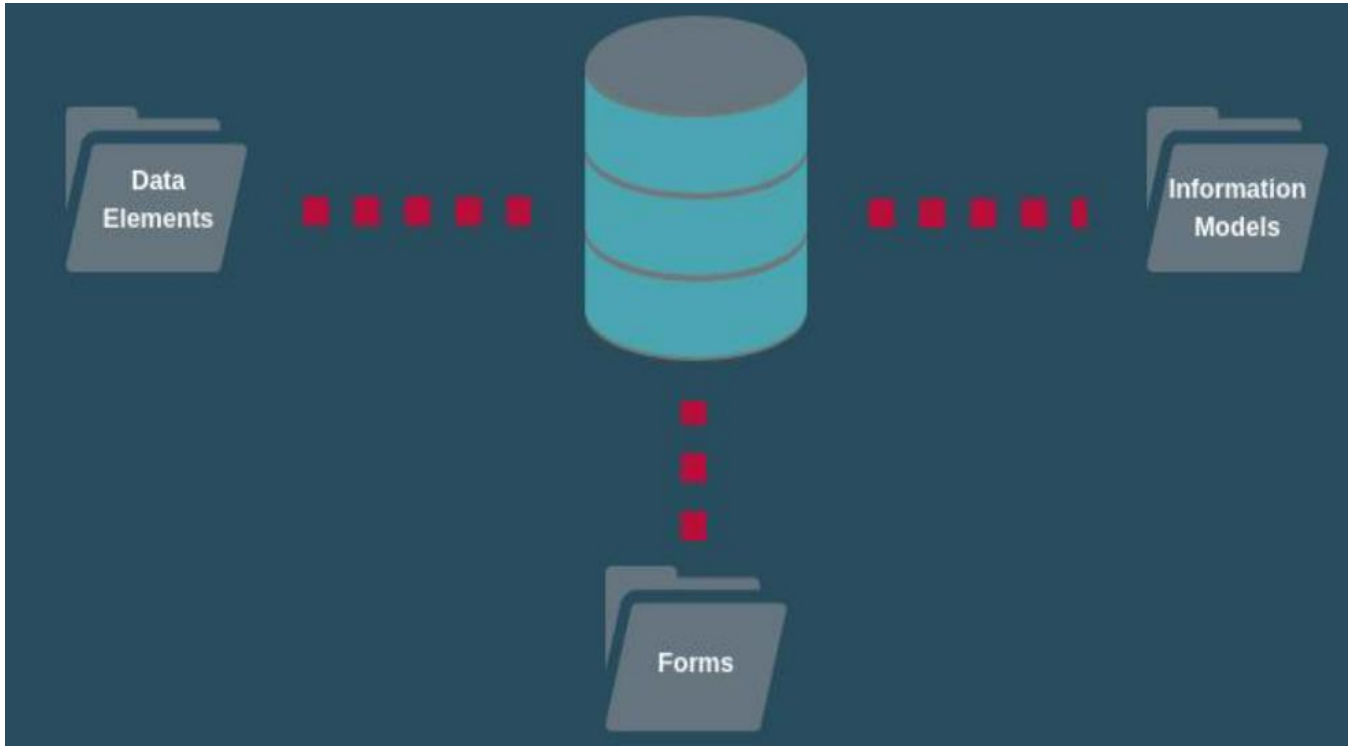


Ontology - caDSR

An **ontology** is a system of carefully defined terminology, connected by logical relationships, and designed for both humans and computers to use.

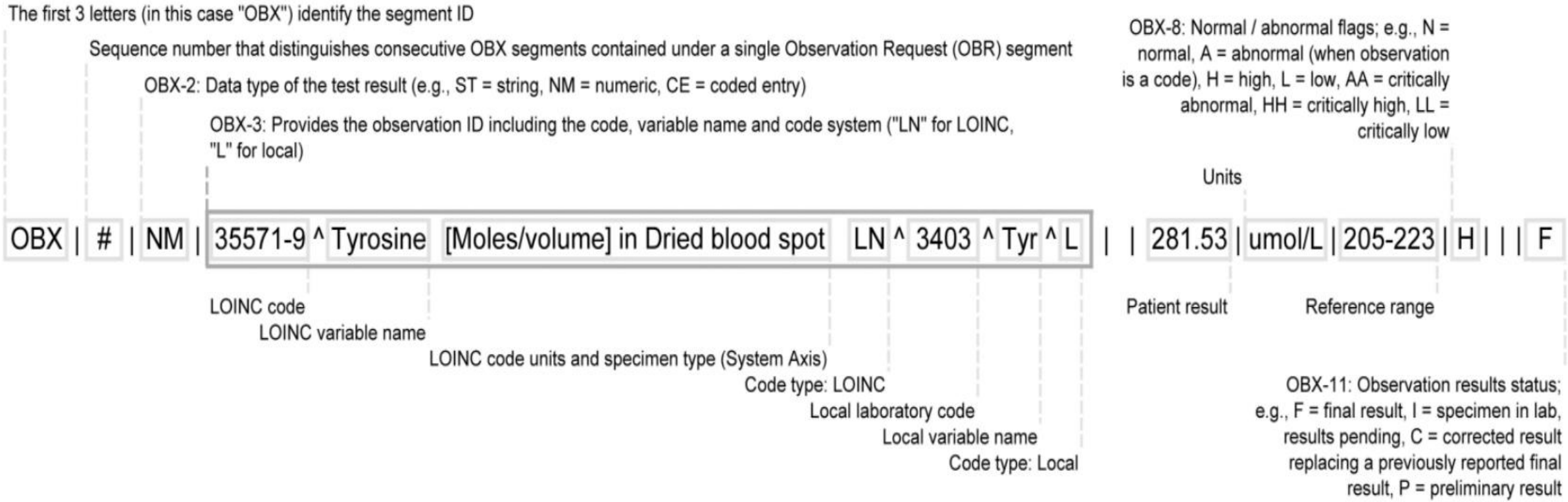
- The GDC project uses the simple ontology **Cancer Data Standards Registry and Repository (caDSR**) developed by the National Cancer Institute (NCI), which builds upon the common data elements (CDEs) to define data and metadata.

- Research protocols use many **CRFs** to collect the data researchers are studying. An **information model** is a software engineering representation of the concepts about cancer research and clinical care.



Ontology - LOINC

- The Logical Observation Identifiers Names and Codes (LOINC) system is supposed to facilitate **interoperability**, and it is the federally required code for exchanging laboratory data.



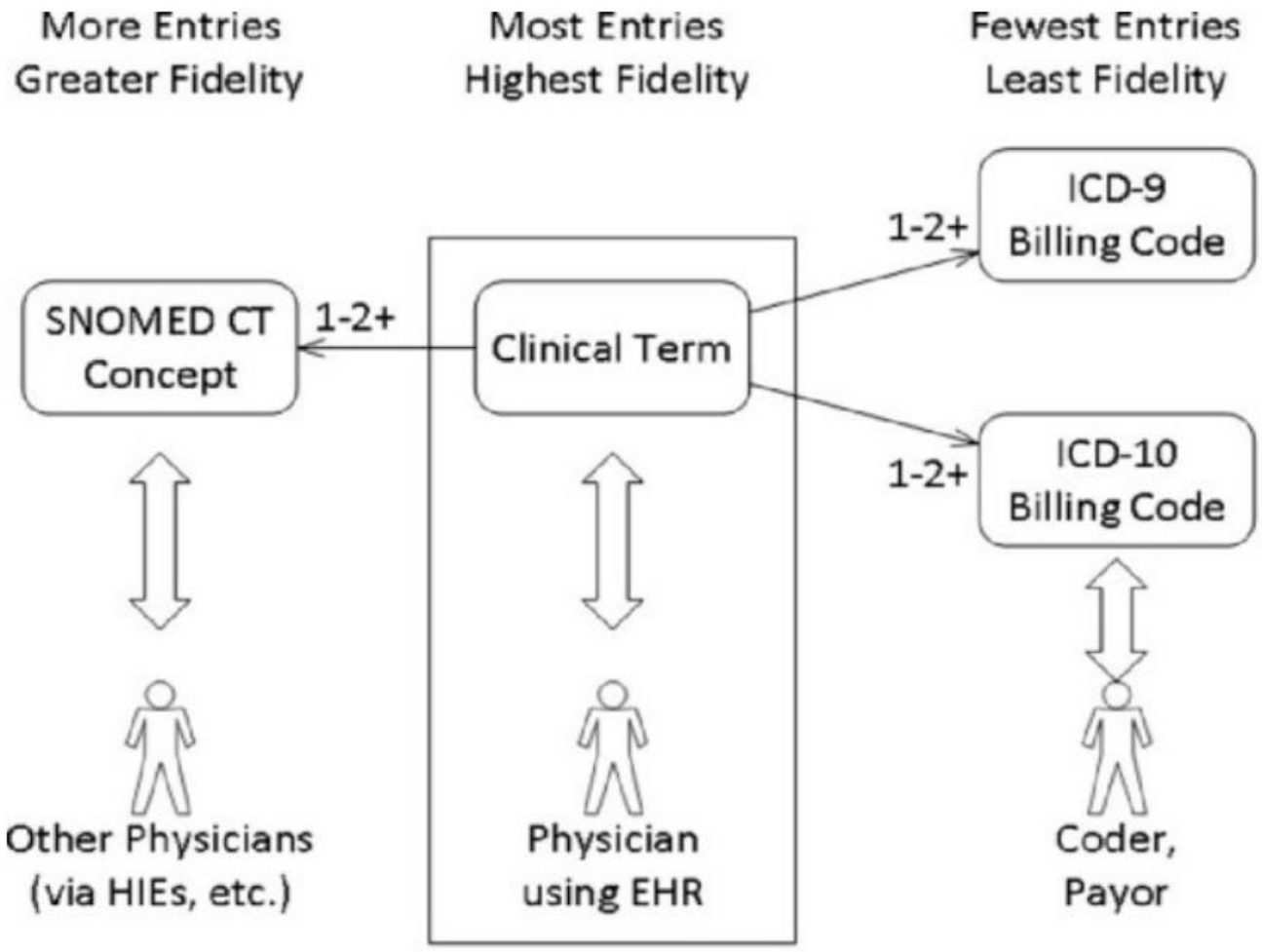
Sample Health Level Seven (HL7) message for newborn screening. Michelle Stram, 2019

Ontology - SNOMED CT

Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT)

- The SNOMED CT is a systematically organized computer-processable collection of **medical terms** used in clinical documentation and reporting.

- SNOMED CT is considered to be the most comprehensive, multilingual clinical **healthcare terminology** in the world. It is now a globally accepted nomenclature and is notably collaborating with **LOINC**.



Duwayne Willett, 2018

Classification – ICD, ICD-O

- **Classifications** are similar to **ontologies** as they define a common language, but they are much narrower in terms of scope. Moreover, there has been a much stronger movement towards convergence regarding classifications than ontologies. Most of this convergence was made possible through the World Health Organization (WHO).
- The International Classification of Diseases (**ICD**) is designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics.
- The ICD for Oncology (**ICD-O**) is a domain-specific extension of the International Statistical Classification of Diseases and Related Health Problems for **tumor diseases**

World Health Organization, 2019

Classification - ATC

Anatomical Therapeutic Chemical (ATC)

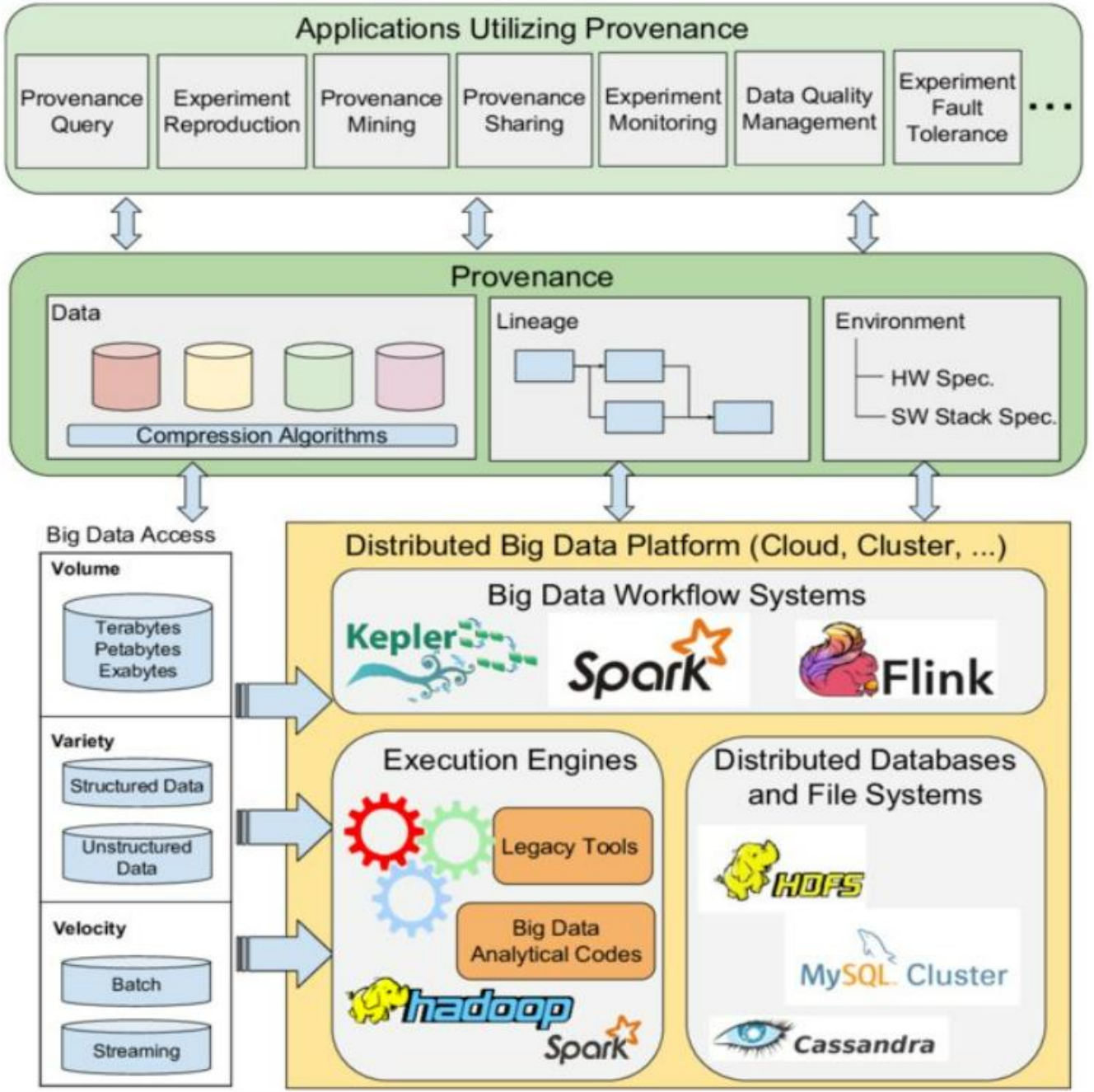
- The ATC Classification System is a **drug classification** system that classifies the active ingredients of drugs according to the organ or system and their therapeutic and chemical properties.
- Its purpose is an aid to monitor drug use and for research to improve quality **medication** use.

Symbol	Description
A	Alimentary Tract And Metabolism
B	Blood And Blood Forming Organs
C	Cardiovascular System
D	Dermatological
G	Genito Urinary System And Sex Hormones
H	Systemic Hormonal Preparations Excl. Sex Hormones
J	General Anti-Infectives For Systemic Use
L	Anti-Neoplastic & Immunomodulating Agents
M	Musculo-Skeletal System
N	Nervous System
P	Anti-Parasitic Products
R	Respiratory System
S	Sensory Organs
V	Various
U	Other
X	Total

Lisa Nissen, 2005

Provenance Information

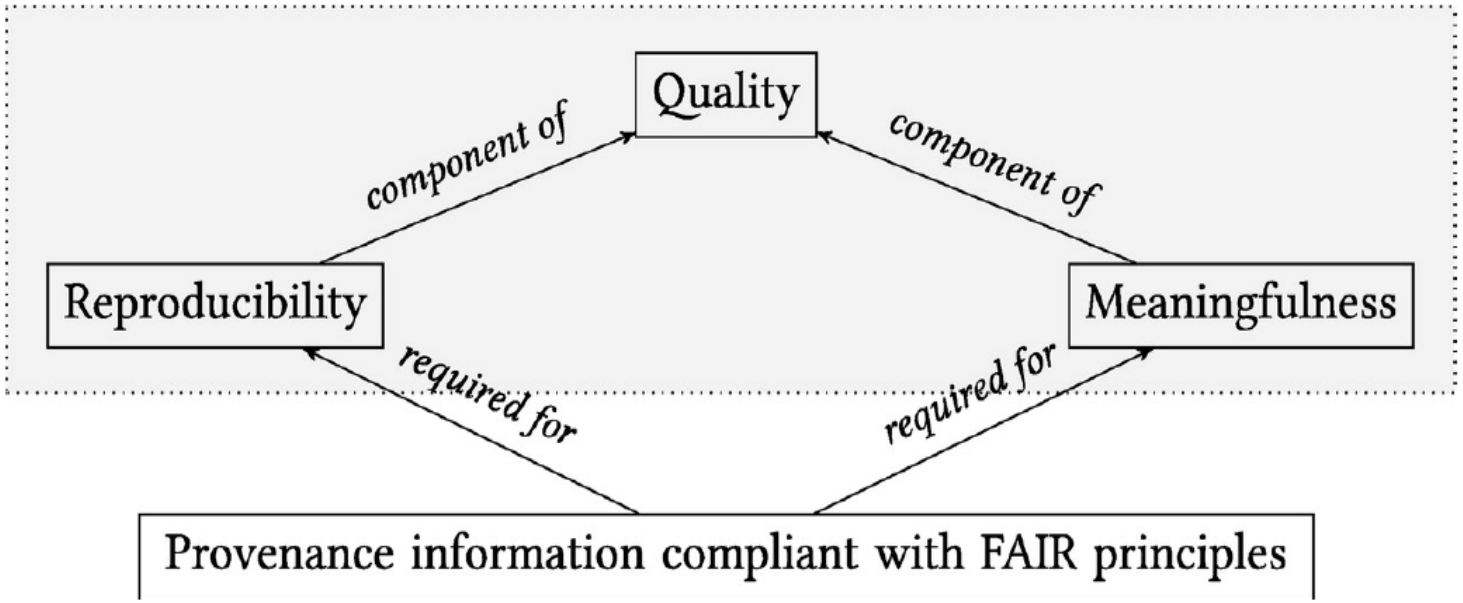
- **Provenance** is information about entities, activities, and people involved in producing a piece of data or thing. FAIR Principles are likely to address only **parts of the problem**.
- Complete provenance information should view **data, lineage, and environment** as a unified resource for reproducibility and meaningful integration of the data.



Jianwu Wang, 2015

FAIR-Health Principles

- An extension of the FAIR Principles was proposed to **FAIR-Health principles**, which include additional quality aspects related to research reproducibility and meaningful reuse of the data.



Petr Holub, Biopreservation and Biobanking, 2018

Relationship between provenance information and components of quality for biological material and data

Genomic Data

- Contrary to clinical data that can be stored and shared directly upon collection, a **genomic data** analysis starts with **biospecimens** of various origins (biopsy, blood, bone marrow, etc.) from which DNA or RNA is extracted.
- The genomic data are then generated from this material and often require further **bioinformatics processing** before it can be interpreted. The entire work-flow needs to be standardized and documented to guarantee **interoperability** and **reusability**.
- More samples may not be available: data is all that is left!
- More details in the upcoming Best Practices lecture!

MOFFITT CANCER CENTER

Charles Vesteghem, et al, Briefings in Bioinformatics, 2020

Standardized File Formats – Critical for FAIR Analyses

- FASTQ file: text file of raw sequence data and base quality scores: https://en.wikipedia.org/wiki/FASTQ_format
- SAM/BAM file: text and binary representation of sequence alignment to a reference genome. Indexed for fast access anywhere in a large file. Well documented: https://samtools.github.io/hts-specs/SAMv1.pdf
- VCF file: Variant Call Format. Text file containing genetic variants in a nested data structure. Many tools exist to handle analysis of this file type. https://samtools.github.io/hts-specs/VCFv4.2.pdf
- MAF file: Mutation Annotation Format. Another format containing genetic variants that is easier for humans to read as it is a tabular file format. Tools exist to analyze/summarize these files as well. Used by TCGA.

- Text files usually compressed with gzip, and tools available to index text file for rapid access: http://www.htslib.org/doc/tabix.html

MOFFITT CANCER CENTER

Considerations for Sharing Molecular Data

- Many computational analyses use molecular data, including DNA and RNA sequencing, proteomics, metabolomics, expression microarrays, etc. These data are now required to be shared by NIH when included in a grant!
- Genome Data Sharing Required by NIH: Genomic Data Sharing (GDS) Policy NOT-OD-14-124

  https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/
- For human subjects research, patients must be consented for sharing, and Human Subjects regulations must be followed (ie, protection of PHI)

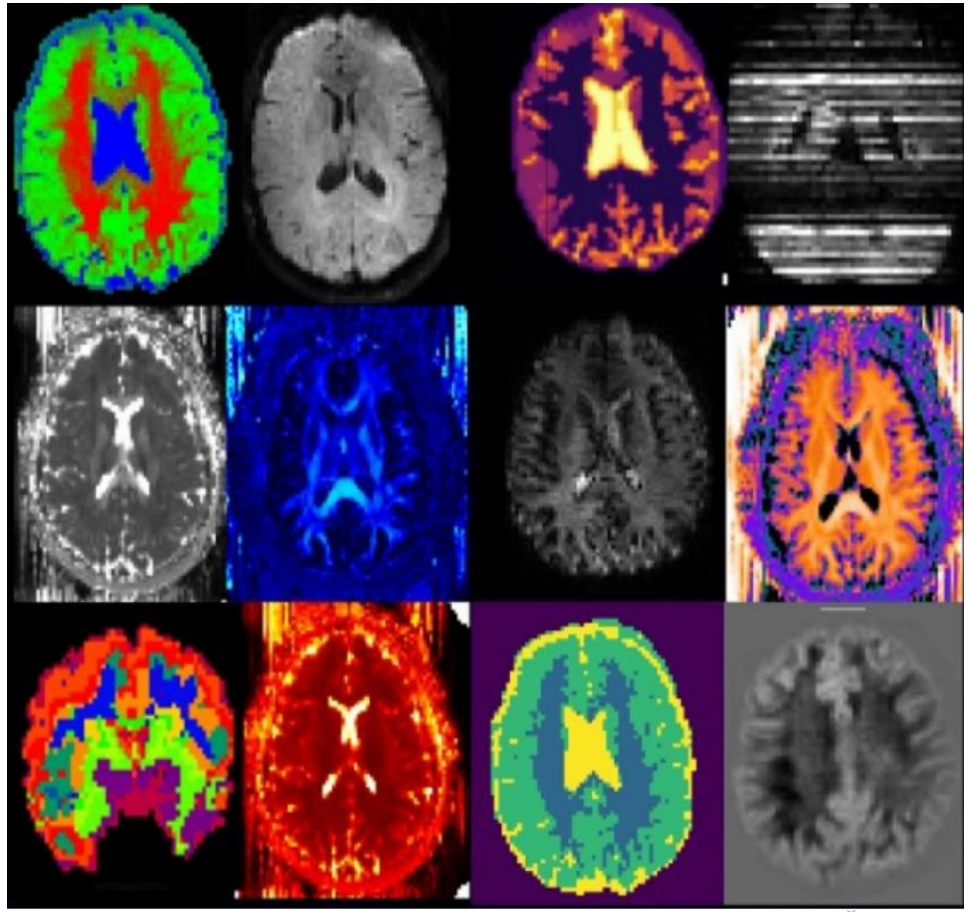
  The Total Cancer Care consent protocol is compatible.
- Moffitt is revising the data sharing process: contact BBSR or Susan Sharpe if you plan to share molecular data.
- PLAN AHEAD!! The process can take a long time!

MOFFITT CANCER CENTER

Imaging Data

- Imaging biomarkers hold tremendous promise for **precision medicine clinical applications**. Development of such biomarkers relies heavily on **image post-processing tools** for automated image quantitation.

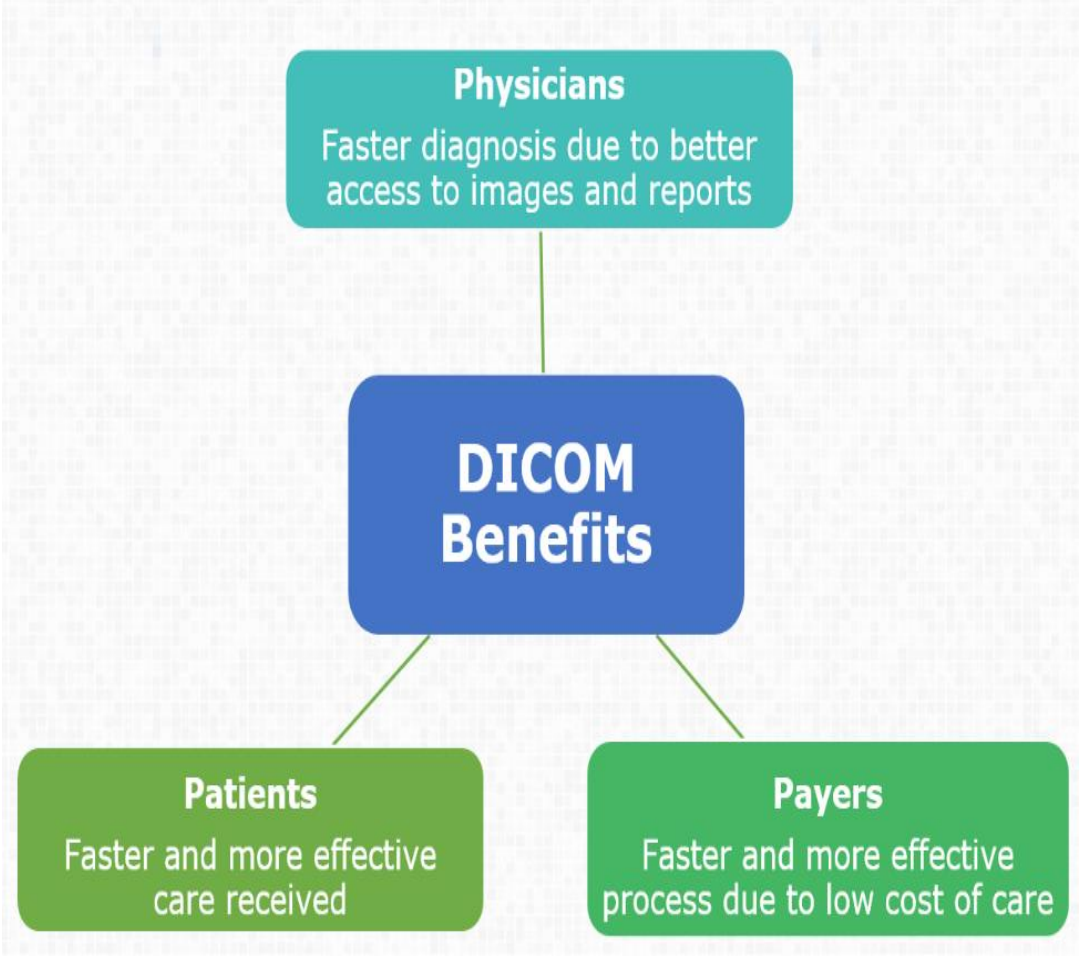
- Their deployment in the context of clinical research requires **interoperability** with the clinical systems.



Jana Hutter, 2020

Digital Imaging and Communications in Medicine (DICOM)

- DICOM is the standard for the **communication** and **management** of medical imaging information and related data. It is most commonly used for **storing** and **transmitting** medical images enabling the integration of medical imaging devices.
- DICOM is primarily used to support **interoperability** between clinical systems for image interchange. Consumption of the DICOM images is widely supported in research tools.



National Resource Centre for EHR Standards, 2022

Question Three:

How to Implement FAIR Principles for AI/ML Data Sharing in **Precision Oncology**?

a. Initiatives to support interoperability and reusability in *data collection*

i. Clinical Data
ii. Genomic Data
iii. Imaging Data

b. Initiatives to support findability and accessibility in data sharing

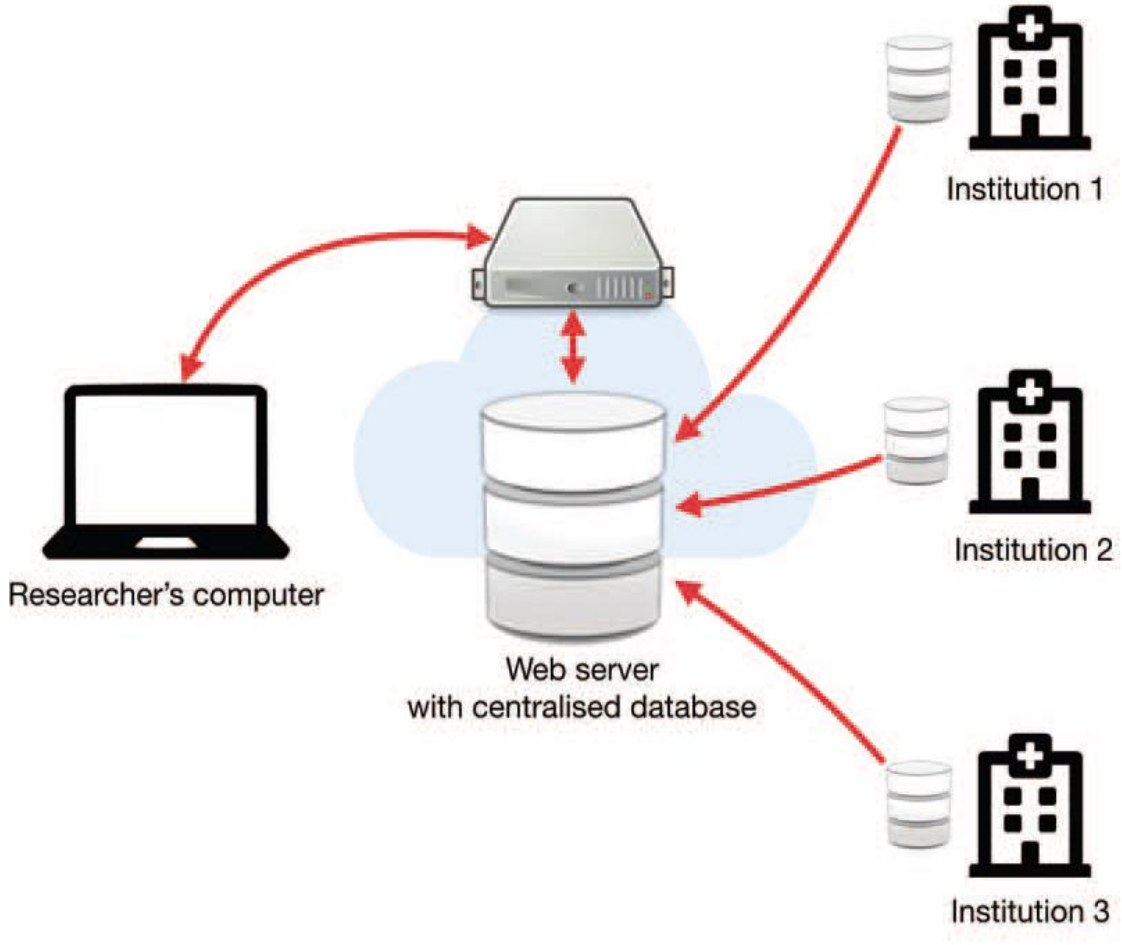
i. Network Architecture
ii. Access Control

MOFFITT CANCER CENTER

Network Architecture - Centralized Architectures

- In the **centralized network architecture**, each institution must upload their data to a **centralized web server**, where everything is gathered in one place.

- This architecture guarantees a better harmonization of the data, but it also faces major **challenges**.
  - The drawbacks of the strict harmonization
  - Very large size of the project
  - The massive quantity of data generated by high-throughput sequencing.



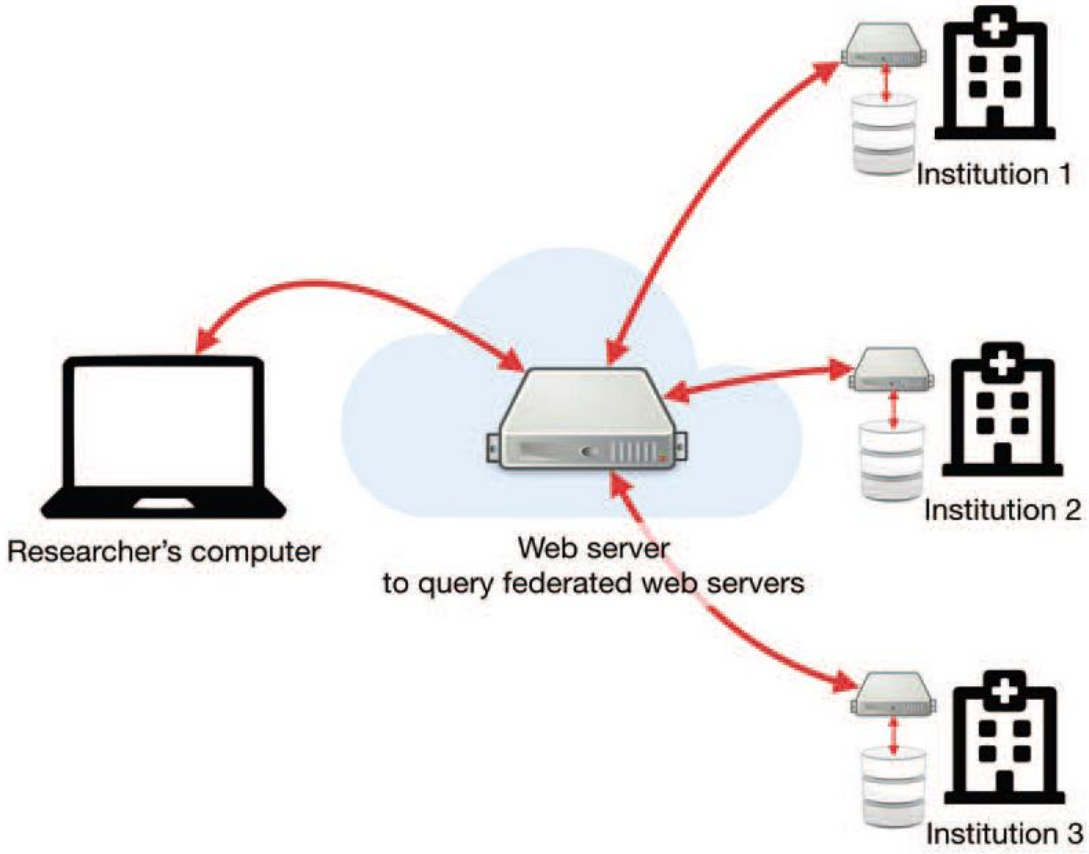
Centralized Architectures

Charles Vesteghem, 2020

Network Architecture - Federated Architectures

- In the **federated network architecture**, data stay at their respective institutions, but each institution must implement an interface to make the data **findable** but not necessarily **accessible.**

- Here, the goal is to make data easily searchable by defining an **interface** rather than a structure. However, the trade-off is that the stored data are not likely to be as **interoperable** as in the centralized scenario.



Federated Architectures

Charles Vesteghem, 2020

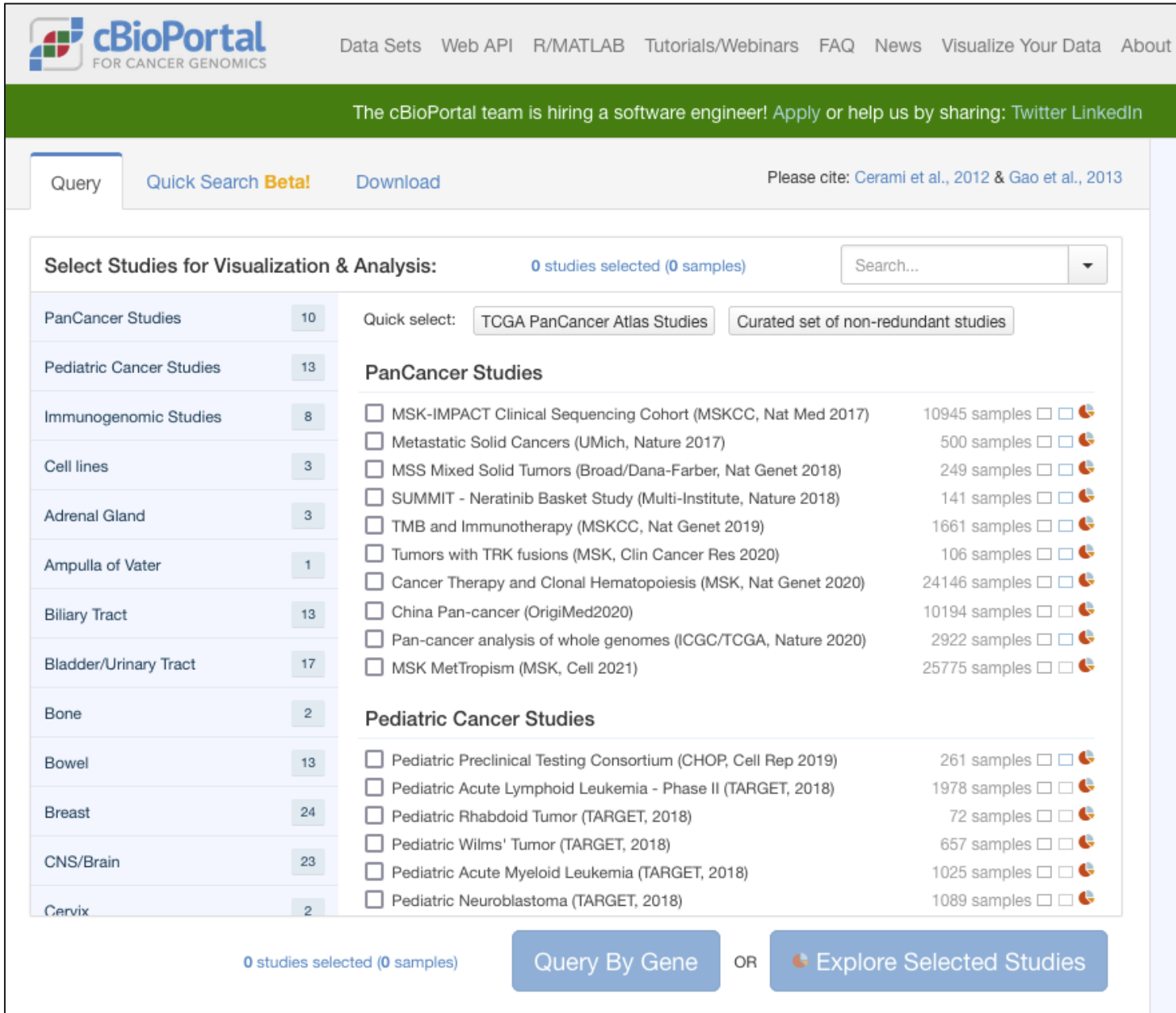
Access Control: Gate-Keeper Vs. Open Access

There are mainly two approaches for accessibility:

- In the **gate-keeper** approach, data are not directly accessible and a request to access data is required. This approach usually guarantees data of better quality and improves the FAIRness of the stored data, notably **reusability**.
- In contrast, the **open access** approach implies that data are available without restriction and its goal is to build common genetic resources to foster research. The main aim is accessibility, potentially to the detriment of other FAIR aspects.
- **There is a trend to have a mix of the two approaches**. The mixed approach followed by the GDC and ICGC seems to be the most pragmatic one, which allows one to keep more sensitive data under control while making less sensitive data easily accessible.

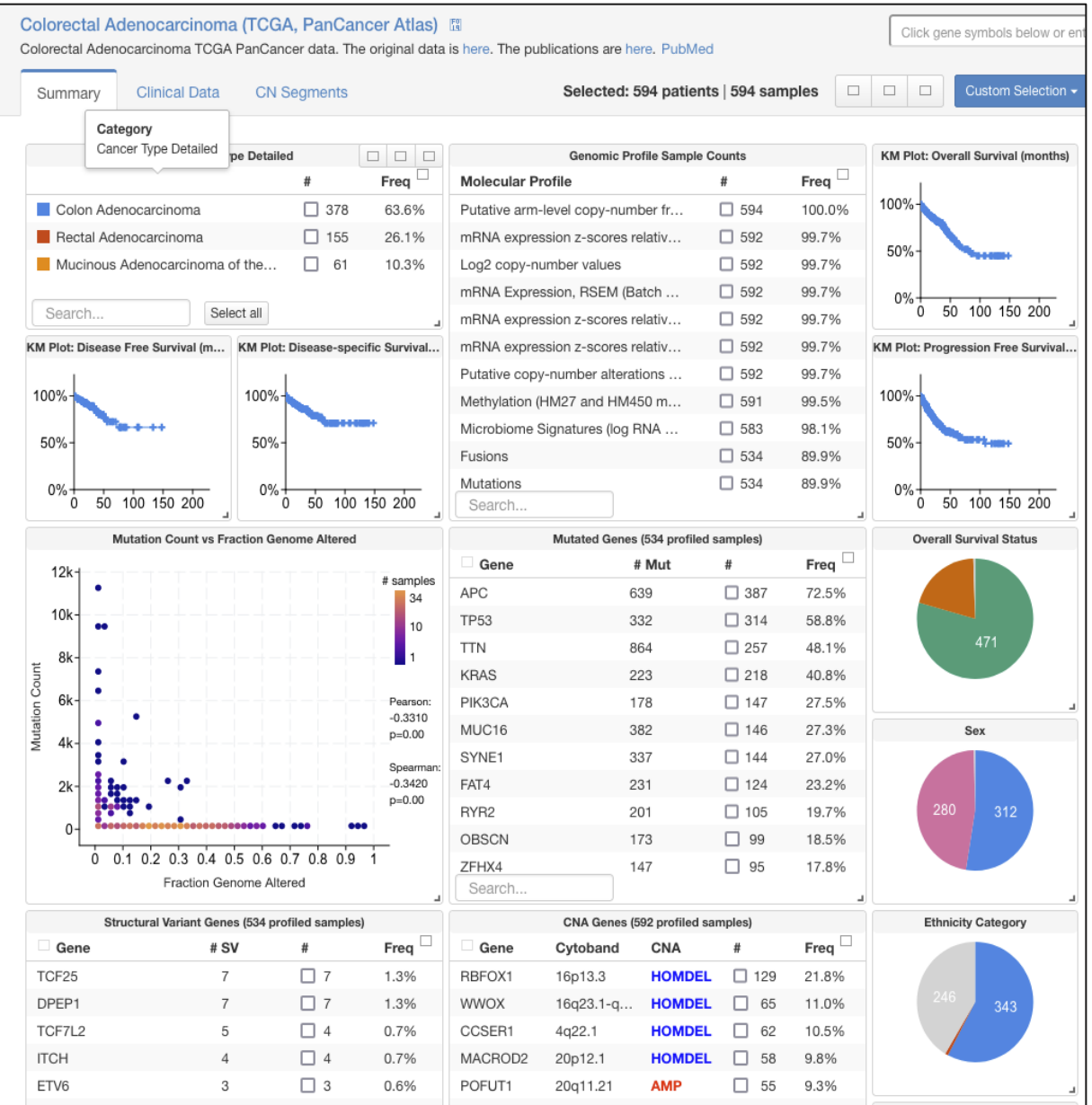
Findable and Accessible Sequence Data - cBioPortal

- Many sequencing projects share high-level results via a web-based tool.
- Includes The Cancer Genome Atlas (TCGA) and others.
- Can easily **find** different diseases.
- Can **access** clinical and molecular results.

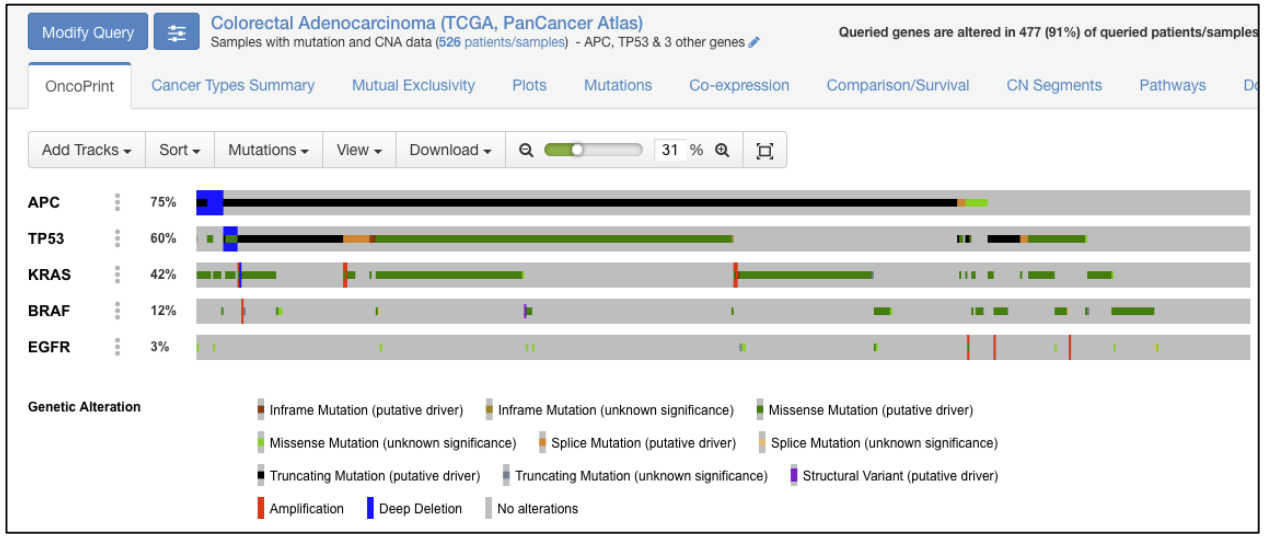


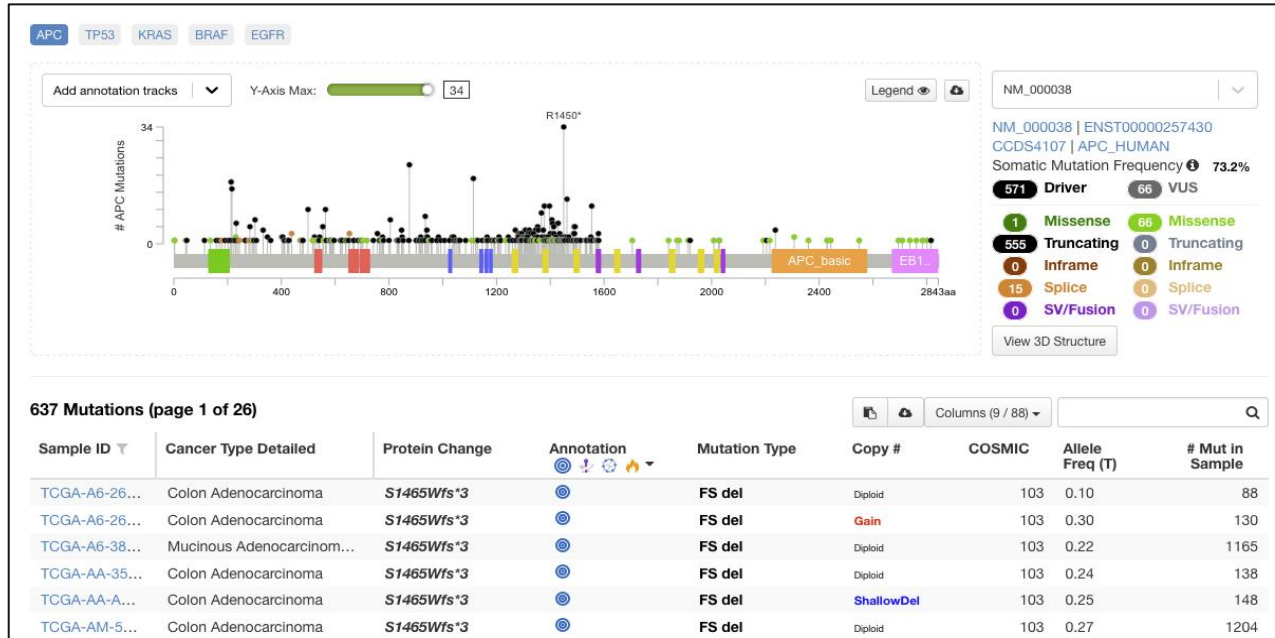
cBioPortal, Colorectal Cancer Example

Clinical Data



Sequence Results (Mutations)





Take Home Messages

- The FAIR Data Principles must be taken into account in the **conception phase** of the project.
- For the actual data collection, **REDCap** is an excellent resource due to its flexibility and open API, allowing it to be easily integrated with existing solutions.
- The **GDC** data structure can be considered the de facto standard and therefore a logical choice for structuring data collection.
- The implementation of existing standards, notably from **WHO**, is mandatory for **interoperability.**
- **Findability** can be achieved through a federated infrastructure.

Questions?

Yi.Luo@moffitt.org
Jamie.Teer@moffitt.org

MOFFITT CANCER CENTER