Study Unit 8

FAIR Data for Health

FAIR Data for Health Outline

- FAIR Data Points and its role in research and medicine
- FAIR Data Principles in Research and Healthcare
- FAIR Data Trains

Study Unit Duration

This Study Session requires a 2 hours of formal study time.

You may spend an additional 2-3 hours for revision

Preamble

Data-driven technologies are changing business, our daily lives, and the way we conduct research more than ever. In recent years, more and more data have been generated in the healthcare ecosystem. The data contain potential knowledge to transform health care delivery and life sciences. Advanced analytics could potentially power the data collected from numerous sources to improve prevention, diagnosis and treatment of diseases, as well as supporting individuals and societies to maintain their health and well-being. The era of exponential growth of data has also witnessed the increase of risk involved in sharing them

This Study Unit teach the importance of FAIR Data Principles in Healthcare research. How FAIR Data Principles can facilitate knowledge discovery from health data. How linked health data drives research, better use and learning from data, and further contributions to patient care.

Learning Outcomes of Study Session 8

Upon completion of this study unit, you should be able to:

8.1 Describe FAIR Data Points (FDP), its roles, components and benefits to research

8.2 Explain the role and tasks of clinical researcher in relation to FAIR Data

8.3 Describe the Personal Health Train (PHT)

8.4 Explain the components of PHT in relation to FAIR

8.0 Introduction

Research data stewardship refers to the long-term and sustainable care for research data, from study design to data collection, analysis, storage, and sharing. It involves all activities that are required to ensure that digital research data is findable, accessible, interoperable, and reusable (FAIR) in the long term, including data management, archiving, and reuse by third parties. This section is divided into three parts: FAIR Data Points and its role in research and medicine, FAIR Data Principles in Research and Healthcare and FAIR Data Trains

8.1 FAIR Data Points and its role for further research and medicine

Link to the video titled 'Introduction to the FAIR Data Point' for more information Introduction to the FAIR Data Point - YouTube

The FAIR Data Point (FDP) is a software component that allows data owners to expose the metadata of their digital objects in a FAIR manner and allows data users to discover properties about offered datasets (or other types of digital objects) through this metadata. The dataset can, if license conditions allow, also be made publicly accessible. A system is called a **FAIR** Data Point (FDP) because it makes data FAIR; especially with the metadata needed for **F**indability and **R**eusability, and a uniform open way of **A**ccessing the data. The FAIR data point also addresses the **I**nteroperability of the metadata it stores, but it leaves the Interoperability aspects for the data itself to the data provider. FDP uses a REST API for creating, storing and serving FAIR metadata. FDP is a software that, from one side, allows digital objects owners/publishers to expose the metadata of their digital objects in a FAIR manner and, for another side, allows digital objects' consumers to discover information (metadata) about offered digital objects. Commonly, the FDP is used to expose metadata of datasets and other types of digital objects like ontologies, repositories, analysis algorithms, websites, etc.

Table 6.1 Importance of FDP

Importance of FDP
1. FDP is a metadata repository that provides access to metadata in a FAIR way.
2. An FDP ultimately stores metadata (information about data sets).
3. An FDP aims to give anyone the power of putting their own data on the web
4. FDP can be used to describe your data sets in a FAIR way, using standard metadata and make them available through simple WWW protocols

6.1.1 Properties of FDP

A. Distributed Data

The FDP is distributed in nature. Therefore, it accommodates different usage scenarios. For instance, one organization may choose to have one instance of the FDP "centralizing" their metadata offering while other organizations may choose to have different FDP instances, each exposing the metadata of a different department. Moreover, each FDP can expose the metadata of datasets (or other types of digital objects) that are located elsewhere.

B. Data Interoperation

Many different data repositories and datasets should interoperate in order to allow increasingly complex questions to be answered. Data interoperability, however, takes place in different levels, such as syntactical and semantical. A collection of FDPs aims to address some interoperability issues at the metadata level by enabling data owners to share their metadata in a FAIR manner thereby fostering Findability, Accessibility, Interoperability and Reusability. To support metadata interoperability the FAIR Data Point has a configurable metadata schema and the metadata content is stored and exposed in RDF.

6.1.2 Components of FDP

FDP has four main components namely the Metadata Provider, the Data Accessor, the Security Enforcer and the Metrics Gatherer which are described by Table 6.2 and shown by Figure 6.1. A FAIR Data Point can be accessed by a user through its graphical user interface (GUI) and by computational clients through its application program interface (API).

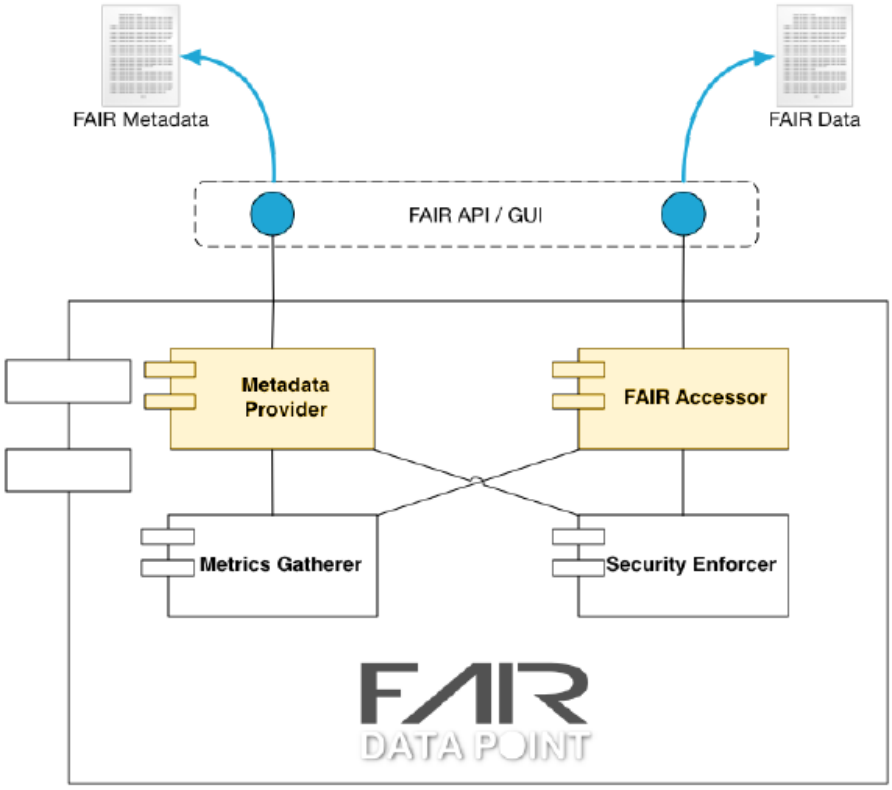


Figure 6.1: FDP Components Source [11]

Table 6.2: Components of FDP

Components	Description
Metadata Provider	This is responsible for giving access to the metadata. The component is accessible as a service to the users through its REST API.
FAIR Accessor	The FAIR Accessor component provides access to the actual data content of the dataset.
Metrics Gatherer	The Metrics Gatherer component monitors various aspects the FDP usage.
Security Enforcer	It acts as a gatekeeper, protecting the access to the (meta)data from requests that do not comply with the given licenses. It reinforces that the "A" in FAIR Data does not necessarily mean open. It means accessible through specific conditions.

6.1.3 Examples of FDP

Several projects that are under way that will use FDP to make data sets known to other researchers:

- Dutch academic hospitals will be implementing FDP To collect COVID-19 data too, with the primary aim of reducing the maintenance burden of several Covid-19 data portals.
- VODAN Covid19 FDP

The VODAN FAIR Data Point (FDP) is developed within the VODAN (Virus Outbreak Data Access Network) implementation network. The VODAN project installs FDP in several different (firstly African) countries, and uses these to collect information on COVID-19 patients. This network will also work on implementing the FAIR Data Train in order to allow distributed analysis of the data.

With a VODAN FDP, organisations across the world can publish their metadata about virus outbreak datasets, such as the WHO's COVID-19 case record form RAPID. Together these FDP's form a network where users can explore metadata informing where data are stored, what definitions and concepts are used and how to gain access to them. The goal is to evolve the VODAN FDP into FAIR Data Stations where data can be accessed and analysed at the station instead of being copied somewhere else. In the VODAN implementation network CODATA, RDA, WDS, and GO FAIR supported by many other organisations work on current challenges around the use and reuse of data pertaining virus outbreaks, like the current COVID-19 epidemic. These challenges range from suboptimal data management, to limited data reuse, lack of semantic interoperability and limited access. The organisations involved in the VODAN implementation network have collaborated to implement the VODAN FAIR Data Point to address these challenges.

The VODAN-in-a-box comes with:

- A semantic data model for the COVID-19 WHO electronic case record form (eCRF) to provide machine-actionable semantics to the data enhancing their interoperability.
- Mappings between the concepts of the COVID-19 WHO eCRF and commonly used vocabularies to enhance interoperability.

- A user-friendly data-entry tool for COVID-19 case notifications based on Covid-19 WHO eCRF.
- A FDP to expose the metadata of the COVID-19-related data.
- Documentation on how to deploy and use all these components.

6.1.4 Open and FAIR Data

The concept of Open Data is more widespread than that of FAIR Data, so it may be necessary to make a clear distinction between the two. Although they have some similarities, they are not exactly the same nor do they have the same audience. The definition of FAIR Data is included in its own name, since it is the abbreviation of Findable, Accessible, Interoperable and Reusable, only that as it will be seen they are not always available for anyone. On the other hand, Open Data, according to the definition of the Open Data Handbook, are "data that can be used, reused and redistributed freely by **any person**, and that are subject, at most, to the requirement of attribution and to be shared in the same manner in which they appear".

Table 6.3 Comparing Open and FAIR Data

Open data	FAIR Data
It is available to everyone to access, use, and share, without licenses, copyright, or patents	It uses the term "Accessible" to mean accessible by appropriate people, at an appropriate time, in an appropriate way. This means that data can be FAIR when it is private, when it is accessible by a defined group of people, or when it is accessible by everyone (open data). It depends completely on the purpose of the data, where the data currently is in its lifecycle, and the end-usage of the data
An example is an undocumented data dump in an uncurated repository, such as OSF, which is neither findable, nor reuseable, nor interoperable)	An example is a data set that is findable, reuseable, etc., but only accessible within a closed research group

6.1.5 Benefits of FDP for research(ers)

Making research data more FAIR will provide a range of benefits to researchers, research communities, research infrastructure facilities and research organisations alike, including:

- Achieving maximum impact from research.
- Increasing the visibility and citations of research.
- Improving the reproducibility and reliability of research.
- Attracting new partnerships with researchers, business, policy and broader communities.
- Enabling new research questions to be answered.

6.1.6 Metadata

If you have data, you have metadata. Metadata is essential to find, reuse and manage your data, and understand the context of your data and files. With metadata you describe who is the responsible researcher, when, where and why the data was collected, how the research data should be cited, etc. The content and format of metadata is often guided by a specific discipline and/or repository through the use of a metadata standard.

A. What is Metadata?

Metadata are data about data. They play an important role in making your data FAIR. Metadata have to be added continuously to your research data, not just at the beginning or at the end of a project. Metadata can be added manually or automatically, and preferably according to a disciplinary standard. From a FAIR perspective, metadata are more important than your data, because metadata would always be openly available and they link research data and publications in the Internet of FAIR Data and Services. The difference between data and metadata is not ontological, but usage. Some researchers' metadata can be other researchers' data. While data documentation is meant to be read and understood by humans, metadata (which are sometimes a part of the documentation) are primarily meant to be processed by machines. The structure of a metadata are described in Table 6.4 and illustrated in Figure 6.2

Table 6.4 Structure of metadata

Types of metadata	Description
Administrative metadata	They are data about a project or resource that are relevant for managing the project. It relates to the technical source of a digital asset which includes data such as the file type, when and how the asset was created, usage rights and intellectual property, providing information such as the owner of an asset, where and how it can be used, and the duration a digital asset can be used for those allowable purposes under the current license. for example, project/ resource owner, principal investigator, project collaborators, funder, project period, etc. They are usually assigned to the data, before you collect or create them.
Descriptive or citation metadata	They are data about a dataset or resource that allow people to discover and identify it. Examples include authors, title, abstract, keywords, persistent identifier, related publications, file format and dimensions etc.
Structural metadata	They are data about how a dataset or resource came about, how it is internally structured, how a digital asset is organized, whether a particular asset is part of a single collection or multiple collections and facilitates the navigation and presentation of information in an electronic resource. Examples are the unit of analysis, collection method, sampling procedure, sample size, categories, variables, Page numbers, Sections, Chapters, Indexes, Table of contents, how pages in a book are organized to form chapters, or the notes that make up a notebook in Evernote or OneNote etc. Structural metadata have to be gathered by the researchers according to best practice in their research community and will be published together with the data. Descriptive and structural metadata should be added continuously throughout the project.

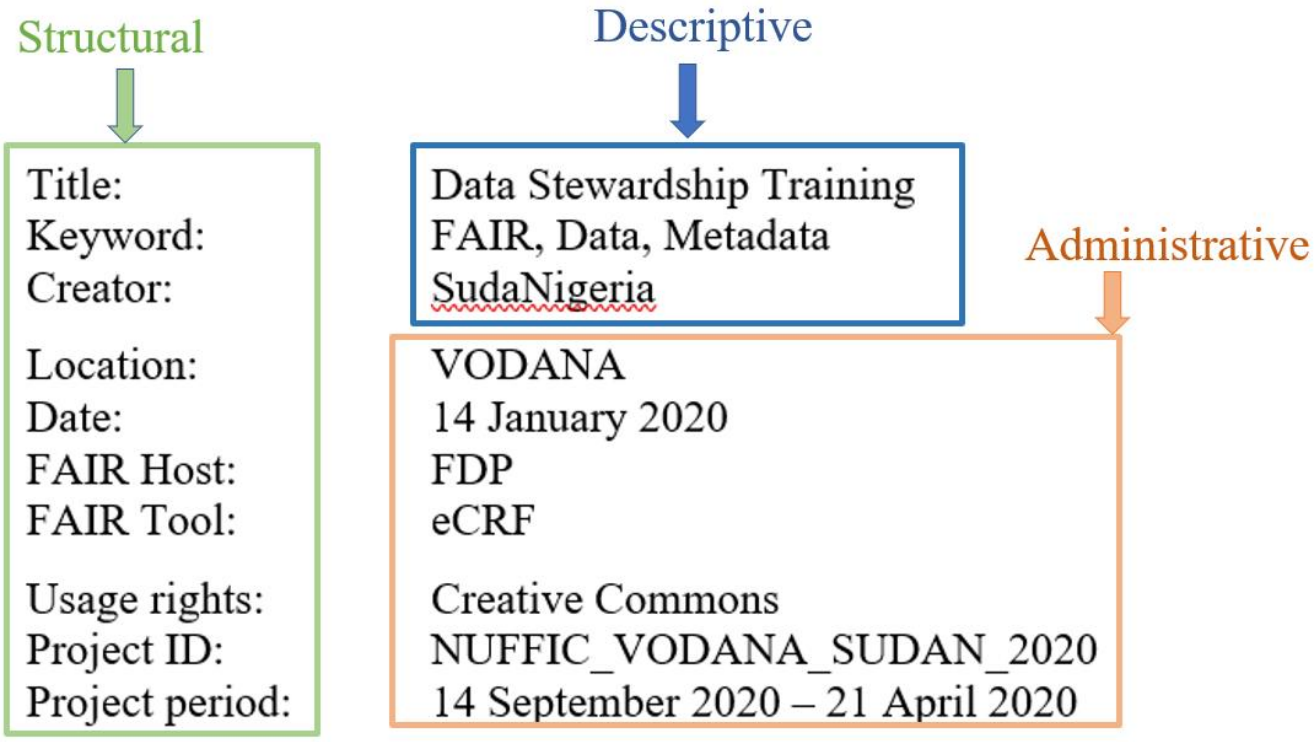


Figure 6.2 Structure of a metadata

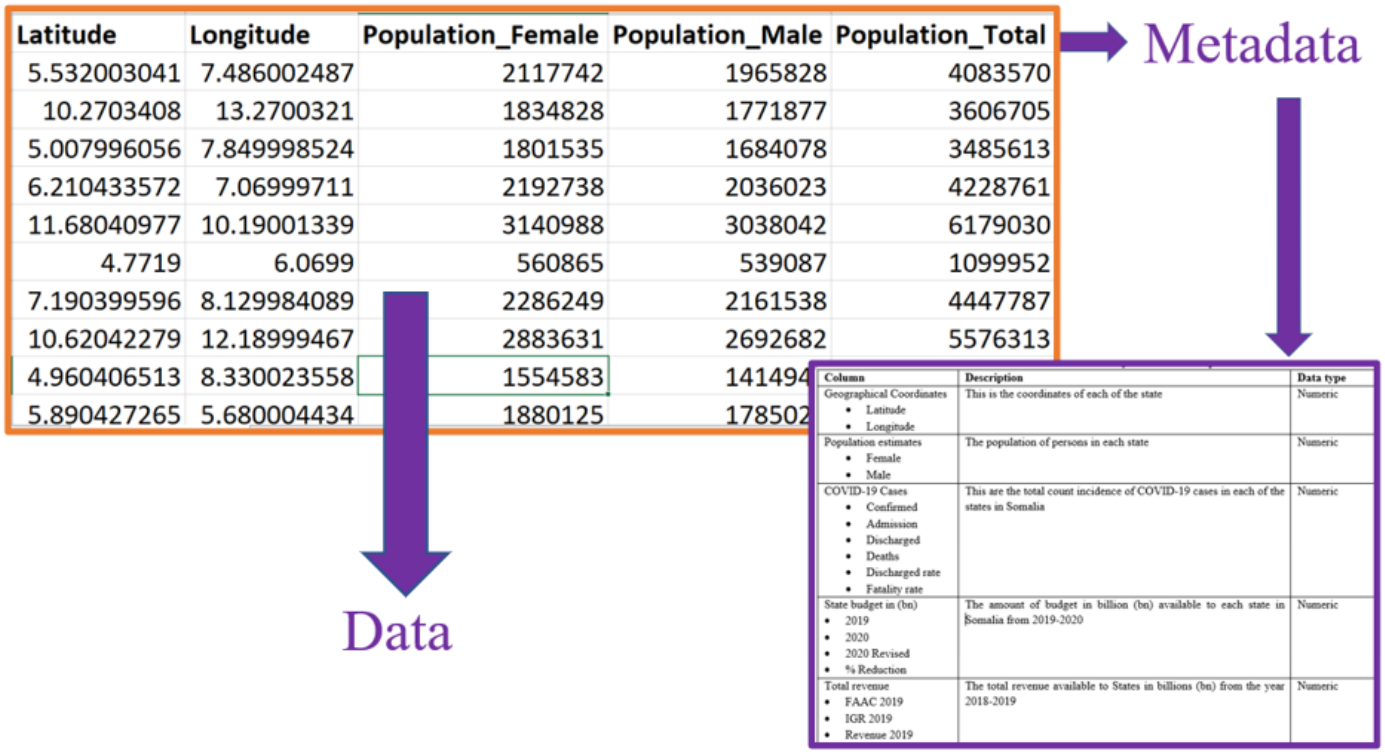
B. Components of metadata

Metadata is an explanation and context of the data. It aids to arrange, find and understand data. A typical metadata should contain the following about the data:

a. Title and description
b. Tags and categories
c. Who created and when?
d. Who last modified and when?
e. Who can access or update?

Find some examples of data and metadata

Example 1 Relational Database

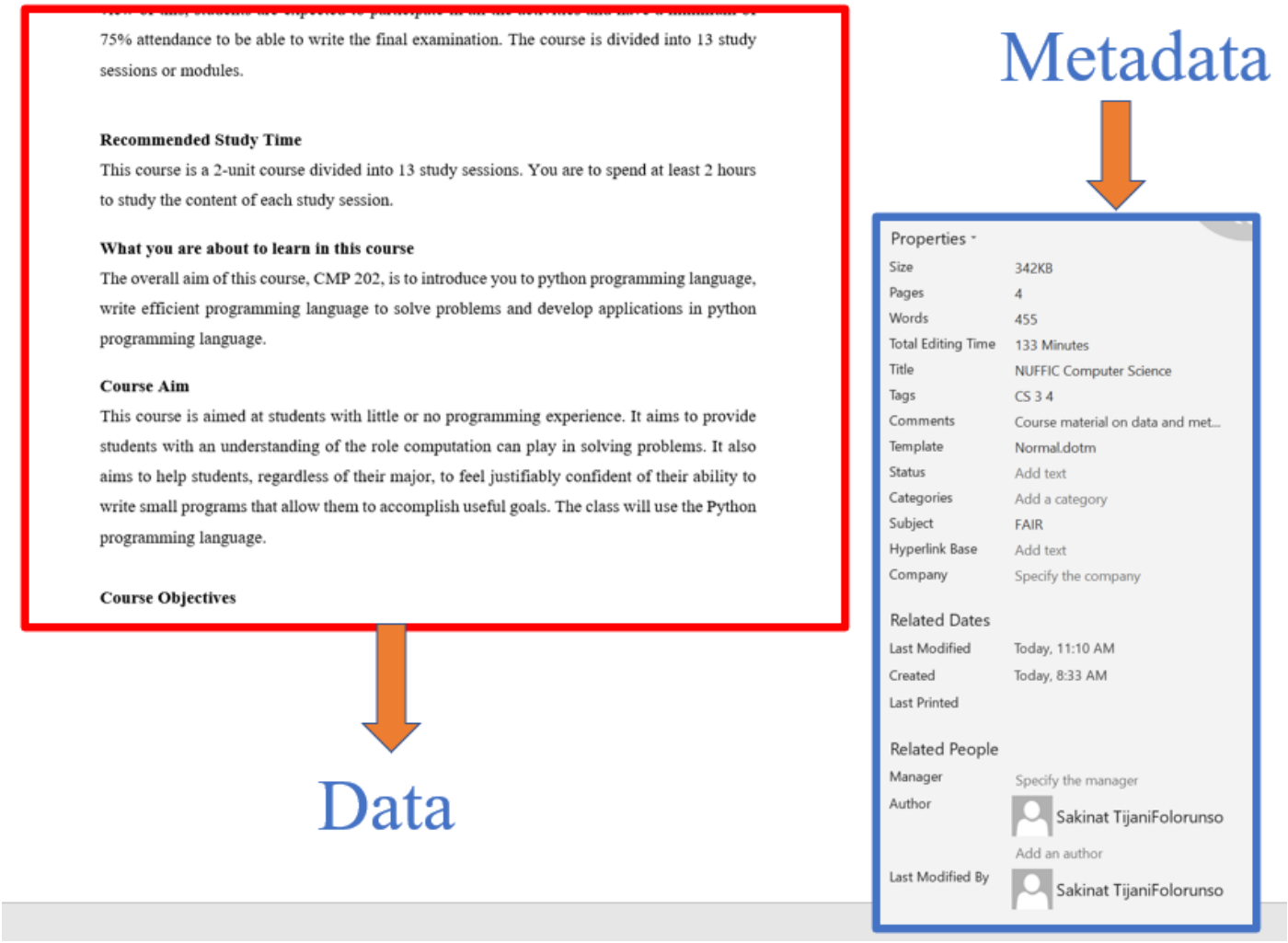


Relational databases stores and give access to both data and metadata. Some of the information it contains are:

- tables,
- rows
- columns,
- data types,
- constraints
- description
- table relationships etc

Example 2 Word Document

All word processing software gathers some standard metadata and enables addition of personal fields for each document.



Some of the typical fields are:

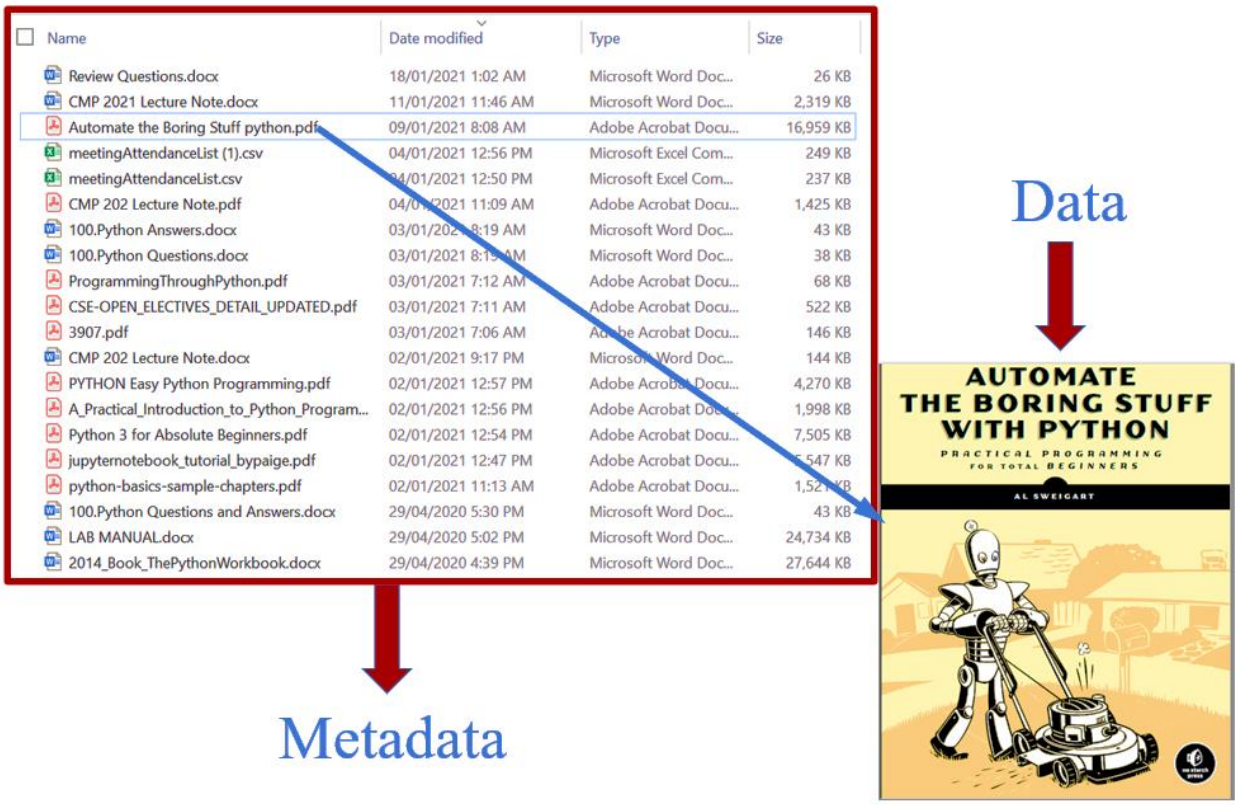
- Title of document
- subject
- author's name
- number of words
- number of pages
- status,

- creation date and time
- who modified last?
- last modification date and time

Example 3 Computer Files

All the fields in individual file explorer is truly metadata. The real data resides inside those files. Some typical Metadata elements includes:

- file name
- file type
- file size
- creation date and time,
- last modification date and time.



C. Metadata standards and ontologies

The quality of metadata is highly important data reusability. It is best practice to use a domain-specific metadata standard and/or an ontology popularly used in your field to describe your data. Some data repositories can help you in choosing the appropriate metadata standard for your data.

C1. What is a metadata standard?

A metadata standard is a subject-specific guide to your metadata. Metadata elements are grouped into sets designed for a specific purpose and given a standard name and definition. Rules on what content must be included, what syntax must be used, or a controlled vocabulary can also be included in a metadata standard. Popular metadata standards for research data are Dublin Core Metadata Standard for bibliographic information, the Data Documentation Initiative (DDI) for survey and observational data (Social Sciences), and the Text Encoding Initiative (TEI) for textual data (Digital Humanities. More resources are available here

So, if there is no standard in a particular field, you joined a team of researchers working on a taxonomy to describe collected data. Metadata standards often start as schemas developed by a particular research group or community to enable the best possible description of their data.

But if the existing metadata standards in your field are incomplete, you can collaborate to define a relevant metadata scheme to suit your purpose.

C2. What is an ontology?

An ontology (or controlled vocabulary) provides a standard definition of key concepts in a particular domain paying attention to how those concepts relates to each other. It can be used to define the structure of your data. An ontology is considered to be a subset of a taxonomy, which is a hierarchically structured conceptual representation of a domain area. Ontologies and taxonomies are closely related, but distinct concepts. For example, the main classes and relations of the African Wildlife ontology (v1) and an illustrative selection of its subclasses as illustrated by Figure 6.3. Two different examples of Healthcare provider ontology at a glance as illustrated by Figures 6.4 and 6.5 and Figure 6.6 illustrating Animal kingdom

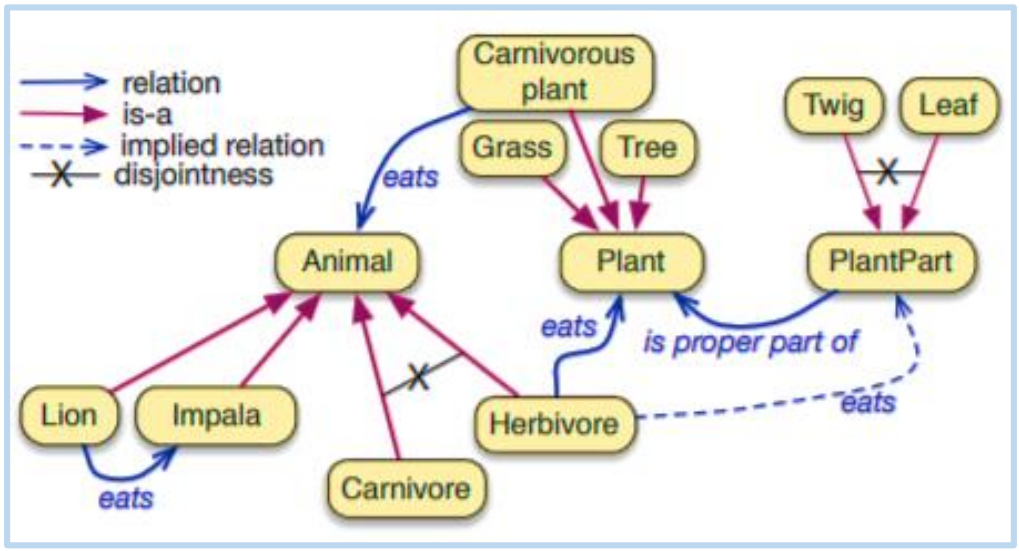


Figure 6.3 The African Wildlife at a glance

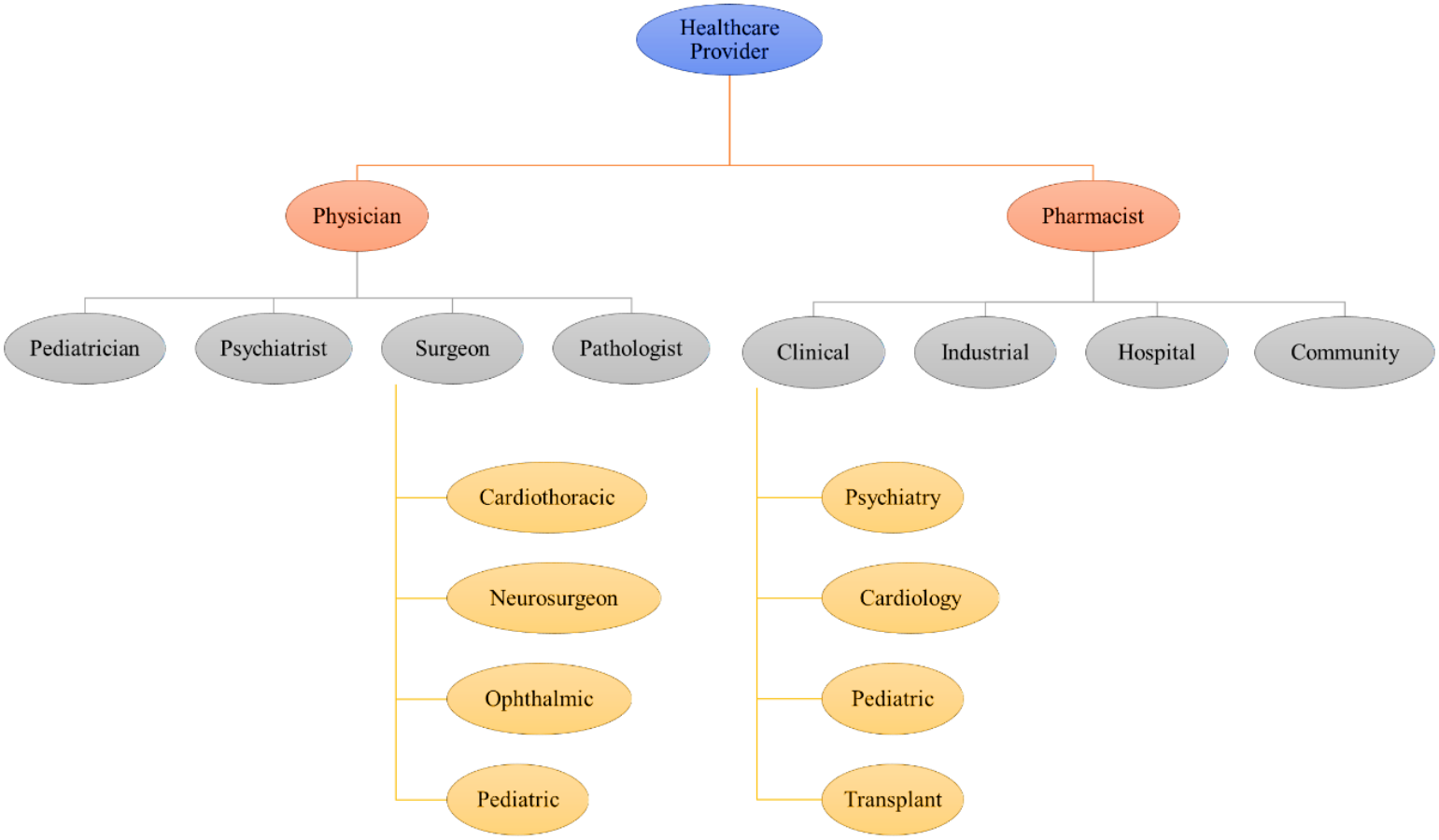


Figure 6.4 The Healthcare provider ontology at a glance

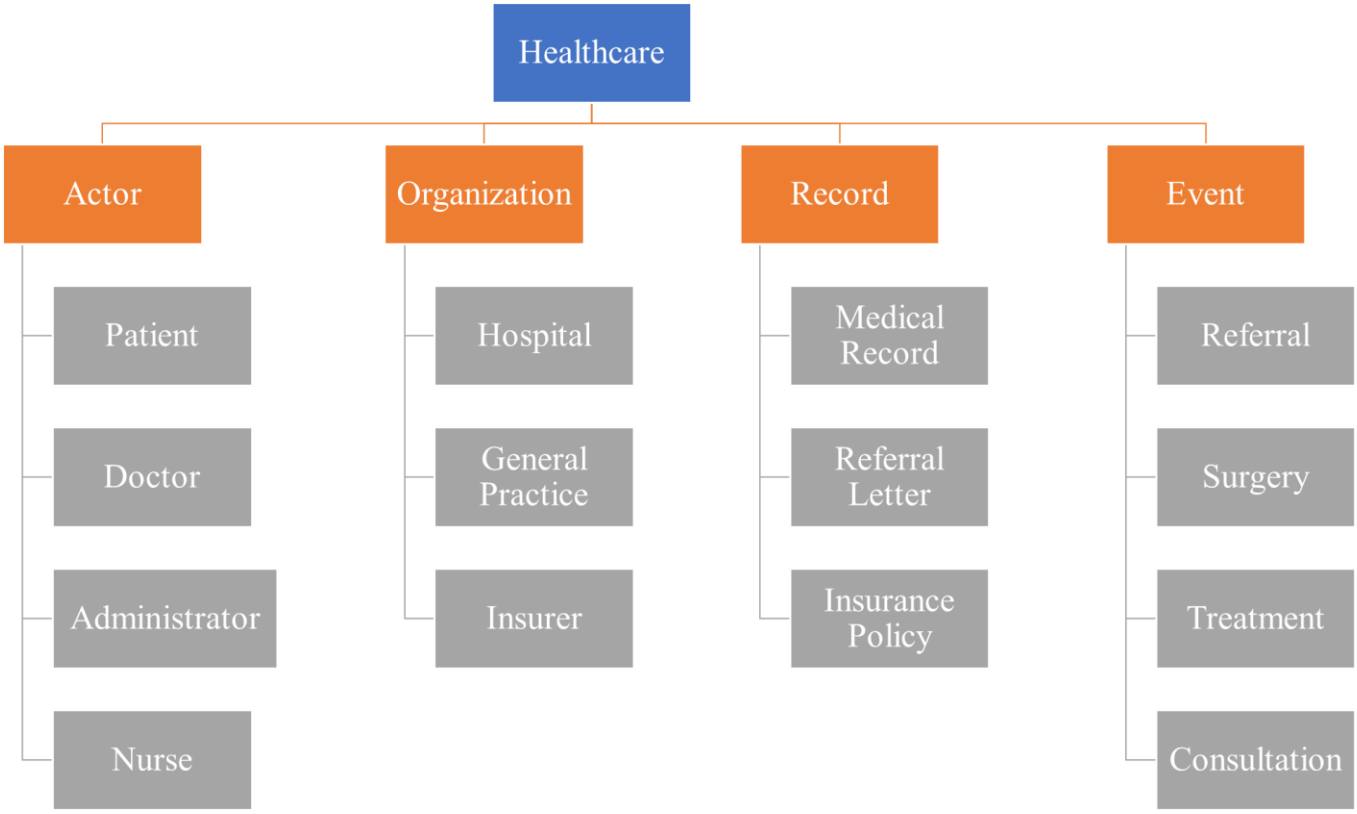


Figure 6.5 The Healthcare provider ontology

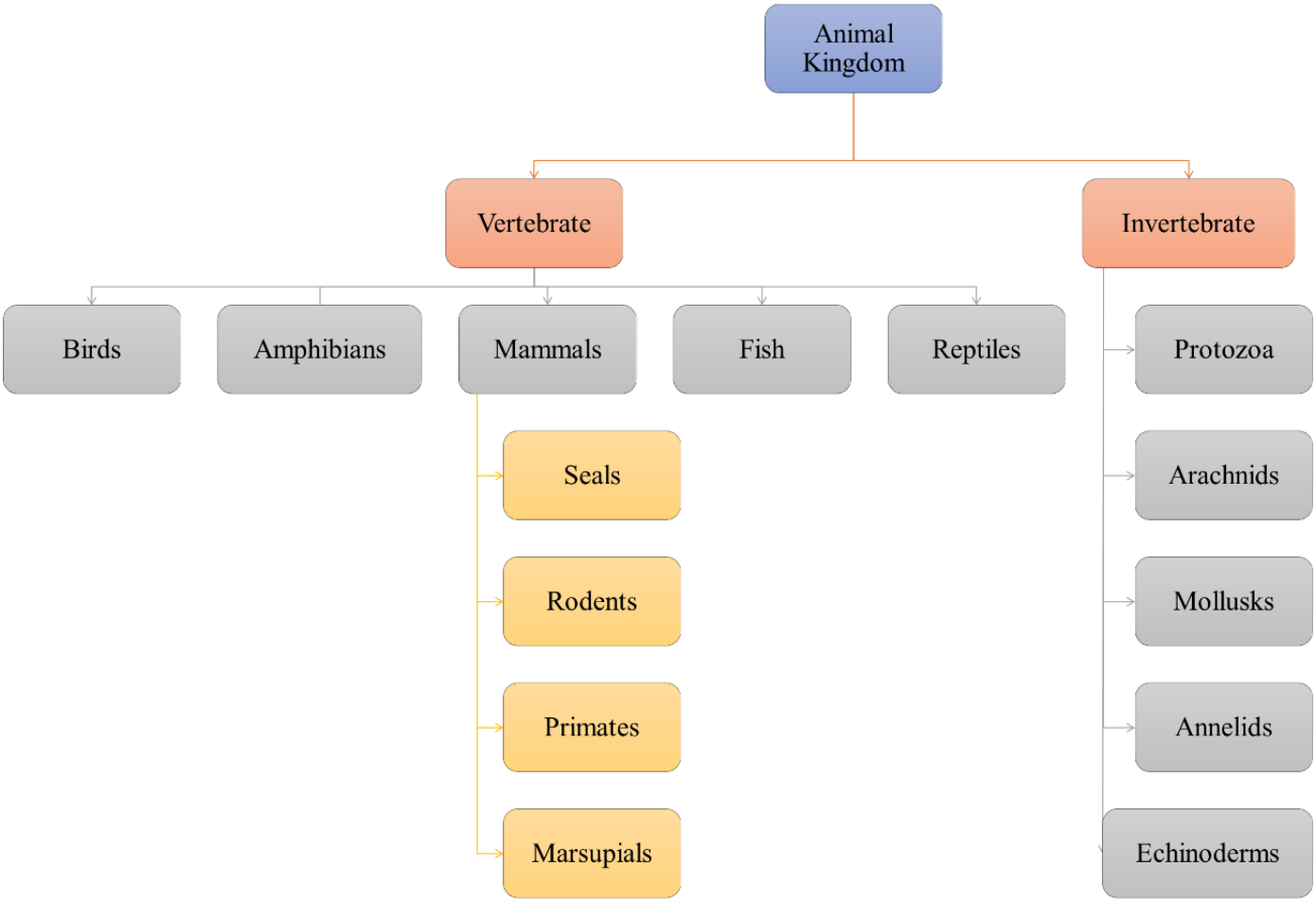


Figure 6.6 The Animal Kingdom ontology

C3 Taxonomy

A taxonomy is a hierarchical categorization of items in a group. It is made up of a tree of categories from the more general at the root of the taxonomy to the more specific at the leaves of the taxonomy. Here is an example of a taxonomy of Healthcare Provider by Figure 6.7: In this example the taxonomy is named "Healthcare Provider" and it consists of categories such as "Physician" and "Neurosurgeon". An item classified in a category in a taxonomy is a member of that category and all of that category's parents. As an illustrative example, imagine an item, "Pediatric",

classified with the category “Surgeon”. It’s also a member of “Physician”, and the “Healthcare Provider” taxonomy as a whole.

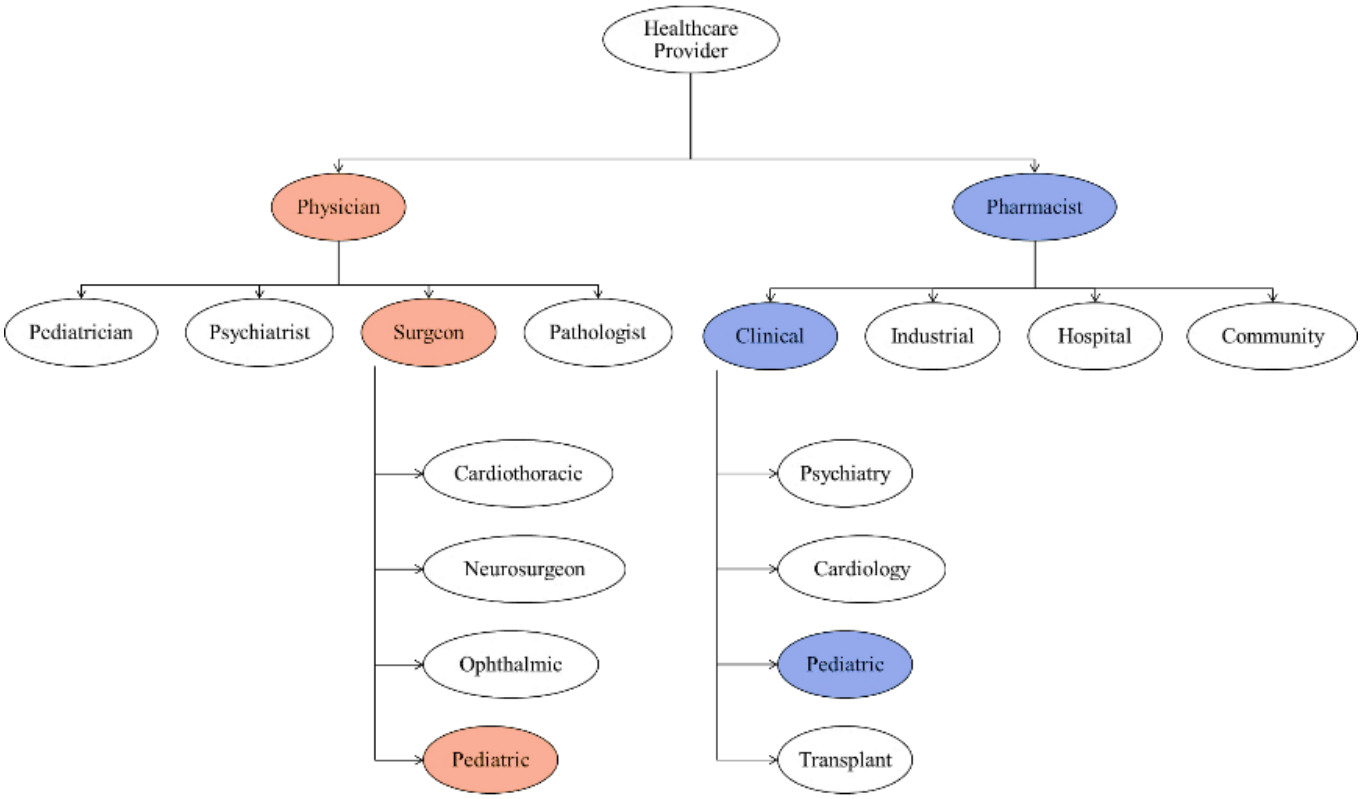


Figure 6.7 Taxonomy of Healthcare Provider

Peer to Peer Interaction

Can you trace the taxonomy of Pediatric?

What other taxonomy can you trace?

C4. Slugs

Each of these resources has a slug that's used when referring to the resource in a URL. A slug is just a unique identifier that's suitable for use in a URL. Standard rules for Slugs in Falkland CMS are:

- alphanumeric (no white space, unicode or special characters)
- lower case
- internally separated by a single dash
- without a prefixed or trailing dash
- 256 characters or less

With Falkland CMS, it is typical for an item to have multiple taxonomies categorizing them.
In addition to being classified in multiple different taxonomies, items can also be categorized in more than 1 category in the same taxonomy. Categories have a slug and a path. The category slug must follow all the standard rules for slugs as stated above in Falkland-CMS and must be unique to all other categories *at the same level and location in the same taxonomy*. The path to a category is made up of the collection, the taxonomy slug, all the parent slugs and finally the category slug. In the healthcare provider taxonomy example, the slug for the "Pediatric" category is "pediatric" and the path is:
/documents/healthcare provider/physician/surgeon/pediatric
Listing items at the path above means the "Pediatric" will be listed since it is classified in the "Surgeon" category, and listing the items in the following paths would also include the "Pediatric":

- /documents/healthcare provider/physician/surgeon/pediatric
- /documents/healthcare provider/physician/surgeon/
- /documents/healthcare provider/physician/
- /documents/healthcare provider/
- /documents/

A fragment example of the JSON representation of the machine learning taxonomy is shown in Table 6.5. More examples of taxonomy are listed here

Table 6.5 showing a fragment example of the JSON representation of the machine learning taxonomy

```
{
  "name":"Healthcare Provider",
  "slug":" Healthcare Provider",
  "collection":"documents",
  "description":" Healthcare Provider",
  "categories": [
        {"physician": "Physician", "categories": [
                {"surgeon": "Surgeon", "categories": [
                        {"cardiothoracic": " Cardiothoracic "},
                        {"neurosurgeon": "Neurosurgeon"},
                        {"ophthalmic": " Ophthalmic"},
                        {"pediatric": " Pediatric"},
                ]},
                {"pathologist": "Pathologist"},
                ]},
          ]
}
```

Using an ontology helps others to understand the structure and content of your data, making your data searchable, interoperable and reusable.

 **Group Work or Project**	Design a taxonomy from any of the following • type of media, type of document • geographic location, topic, time period • biological classification • chemical classification • Dewey Decimal Classification • folk taxonomies

C5. Machine-readable format

Machine-readable format means a structured format that can automatically be read and processed by a computer. Machine readable data is a data in machine readable format. Machine-readable data must be structured data.

Machine-readable data may be classified in two groups: human-readable data that is marked up so that it can also be read by machines (e.g. microformats, RDFa, HTML), and data file formats intended principally for processing by machines (CSV, RDF, XML, JSON). These formats are only machine readable if the data contained within them is formally structured; exporting a CSV file from a badly structured spreadsheet does not meet the definition.

Machine readable is not synonymous with *digitally accessible*. A digitally accessible document may be online, making it easier for humans to access via computers, but its content is much harder to extract, transform, and process via computer programming logic if it is not machine-readable.

Table 6.6 presents some sample machine readable file formats while Table 6.7 shows a Sample RDF, Turtle and JSON-LD file format of the same data. More examples of file format can be found here

Table 6.6 Sample FAIR file formats

FAIR File	Format
Containers	tar, gzip, zip
Databases	xml, csv, json
Geospatial	shp, dbf, geotiff, netcdf
Video	mpeg, avi, mxf, mkv
Sounds	wave, aiff, mp3, mxf, flac
Statistics	dta, por, sas, sav
Images	tiff, jpeg 2000, pdf, png, gif, bmp, svg
Tabular data	csv, txt
Text	xml, pdf/a, html, json, txt, rtf, rdf
Web archive	warc

Additional Resources

More resources on FDP and Metadata are available here

- Metadata Standards Directory external link
- Storing and preserving data
- Learn to write your Data Management Plan external link
- Watch the following video external link explaining structural and descriptive metadata.
- add metadata to your data using Excel external link.

Peer to Peer Interaction

Can you differentiate between open and FAIR data?

What is the importance of FDP?

Is metadata useful?

6.2 FAIR Data Principles in Research and Healthcare

Data stewardship is the long-term, sustainable care for research data. This has become an indispensable part of clinical research. This section describes the aspects of data stewardship that are important in clinical research.

6.2.1 Responsibilities of a clinical researcher

The clinical researcher is the principal data steward. He is responsible for the complete scientific process: from study design to data collection, analysis, storage, sharing and protecting the privacy of study subjects. The formal responsibility for personal data lies with research institutes, which

is accountable for having adequate policies, facilities, and expertise around data stewardship. According to the principle of accountability in the General Data Protection Regulation (GDPR), it is the institute's responsibility to ensure that the fundamental principles relating to processing of personal data are respected, as well as the ability to demonstrate compliance. The research institute should appoint a Data Protection Officer that monitors GDPR compliance at the institute. Table 6.7 provides an overview of the responsibilities of the main people involved in data stewardship for clinical research

Table 6.7: Role of a researcher

Role	Role Description
Researcher	a. Is accountable for research data; b. Is in control of the complete research data flow; c. Reuses existing data when possible; d. Collaborates with patient organisations throughout the research project; e. Protects the privacy and safety of study subjects; f. Applies the FAIR principles; g. Protects research quality and reproducibility; h. Uses available expertise and recommended infrastructure; i. Thinks ahead about intellectual property rights; j. Shares data responsibly
Research institution	a. Employs professionals that provide the procedures and technical systems for data stewardship (e.g., data stewards, data managers, IT-specialists, statisticians); b. Has institute managers, who govern and facilitate the professionals; c. Has supervisory bodies such as medical-ethical review committees and privacy officers; d. Engages with patients and citizens from whom data is collected; e. Offers facilities to protect data according to the GDPR
Manager of research	a. Establishes facilities for data stewardship (e.g., data protection, storage, interoperability);

institution	b. Provides financial means for data stewardship and expert employees; c. Is responsible for organisation, policy, standard procedures, practical measures; d. Ensures training for employees that work with data
Professional that supports data stewardship	a. Provides, gives advice on, and supports the use of terminologies, IT-standards, and e-infrastructure which promote data sharing and integration; b. Gives advice on writing data management sections and plans, metadata standards, repositories, and data handling c. Supports data curation and archiving

Peer to Peer Interaction

Can you describe the different roles of Clinical Researcher?

Can you distinguish the role of institution to an individual researcher?

6.2.2 Major Task of a Clinical Researcher

Some of the major task of a clinical researcher who is also a data steward are preparation of a study, privacy and autonomy, data collection, analysis, archiving and sharing. Their description and specification are described in the following section

Table 6.8: **Task and description of a Clinical Researcher**

TASK	**DESCRIPTION**
Preparing a Study	▪ Study Design and Registration ▪ Re-using Existing Data ▪ Collaborating with Patients ▪ Data Management and Statistical Analysis Plan ▪ Describing the Operational Workflow ▪ Choosing File Formats ▪ Intellectual Property Rights ▪ Data Access
Privacy and Autonomy	▪ Informed Consent ▪ Care and Research Environment ▪ Preparing Sensitive Data for Use
Data Collection	▪ Data Management Infrastructure ▪ Data Monitoring and Validation ▪ Metadata ▪ Security ▪ Access policy ▪ Protecting Research Data
Analyzing Data	▪ Raw Data Preparation ▪ Analysis Plan
Archiving Data	▪ Archiving: What and How? ▪ Archiving: Where?
Sharing Data	▪ General Considerations ▪ Anonymity ▪ Sharing with Commercial Parties

Peer to Peer Interaction

What areas are important in the preparation for a study process?

Describe the basics of Data Management

6.3 FAIR Data Trains

The Personal Health Train (PHT) aims to connect distributed health data and create value by increasing the use of existing health data for citizens, healthcare, and scientific research. The key concept in the PHT is to bring algorithms to the data where they happen to be, rather than bringing all data to a central place. The PHT is designed to give controlled access to heterogeneous data sources, while ensuring privacy protection and maximum engagement of individual patients and citizens. As a prerequisite, health data is made FAIR (Findable, Accessible, Interoperable and Reusable). Stations containing FAIR data may be controlled by individuals, (general) physicians, biobanks, hospitals and public or private data repositories.

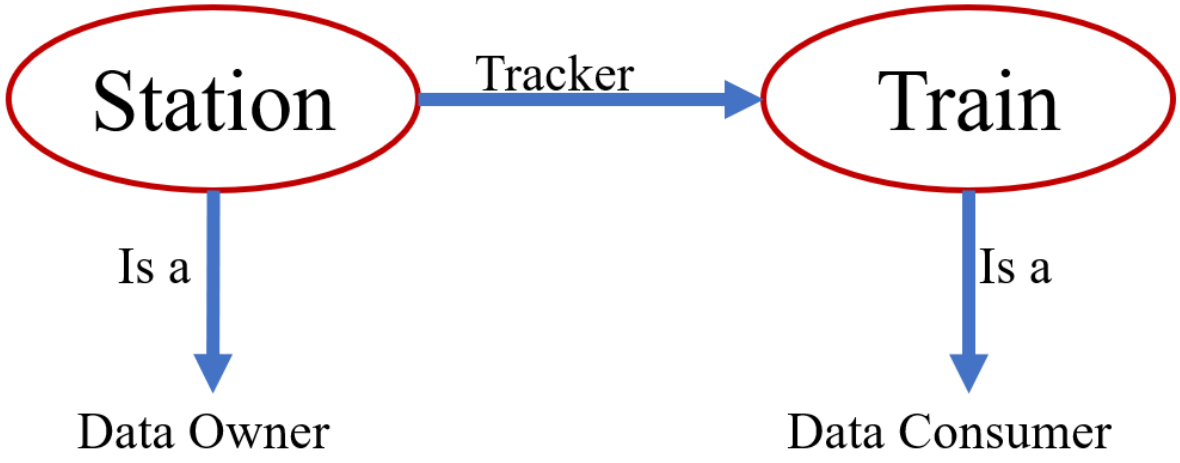


Figure 6.8 Data Train

PHT proposes an alternative approach to existing data sharing and licensing agreement to which encompasses both technological and social aspects of sensitive data reuse. When data sharing is not achievable, using distributed analytics on distributed data becomes a viable solution. The PHT does not require the transfer of data from the holding entity. Rather than moving the data to the requester, it moves the analytics tasks to the data repositories and executes the tasks in a secure environment. In this approach, the owner of the data can remain in control and decide which part of the data will be analysed for which specific purposes and by whom. This new approach requires discovering, understanding, exchanging and executing analytics tasks with minimum human intervention. FAIR principles become relevant not only for data but also for analytics tasks. In the fragmented landscape of data, interoperability and accessibility can be ensured by applying FAIR principles to the analytics tasks and system components that interact with these tasks. In this paper, we will demonstrate the application of FAIR principles to the Personal Health Train approach.

6.3.1 Importance of Personal Health Train (PHT)

PHT is important to ecosystem development and analytic. Some of the major highlights are:

a. The PHT provides an infrastructure to support distributed and federated (AI) solutions that utilize the data at the original location.

b. The PHT does not prescribe any specific standard or technology for data, and instead, it only requires publishing individual choices as metadata.

c. The PHT focuses on making data, tasks, processes and algorithms findable, accessible, interoperable and reusable (FAIR).

d. The PHT provides an alternative solution to reuse the data in institutional data silos or citizens' personal data stores.

e. The PHT approach could unleash the potential of big data analytics for personal data without compromising privacy.

6.3.2 Components of architecture of PHT

The PHT defines the following three core components. They are Station, Train and Tracker or handler. Figures 6.9, 6.10 and 6.11 shows the individual components and their description

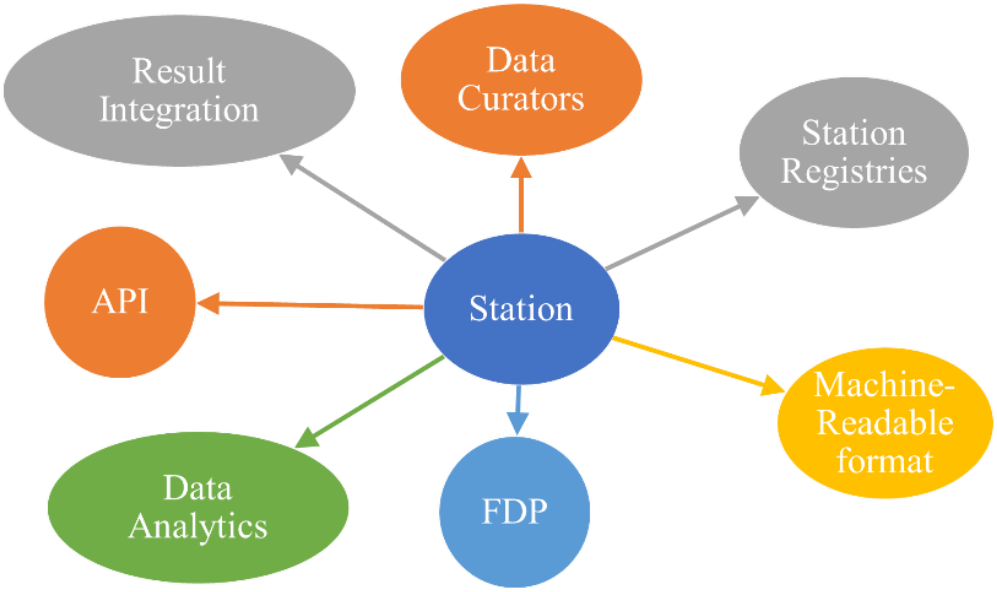


Figure 6.9 The Station

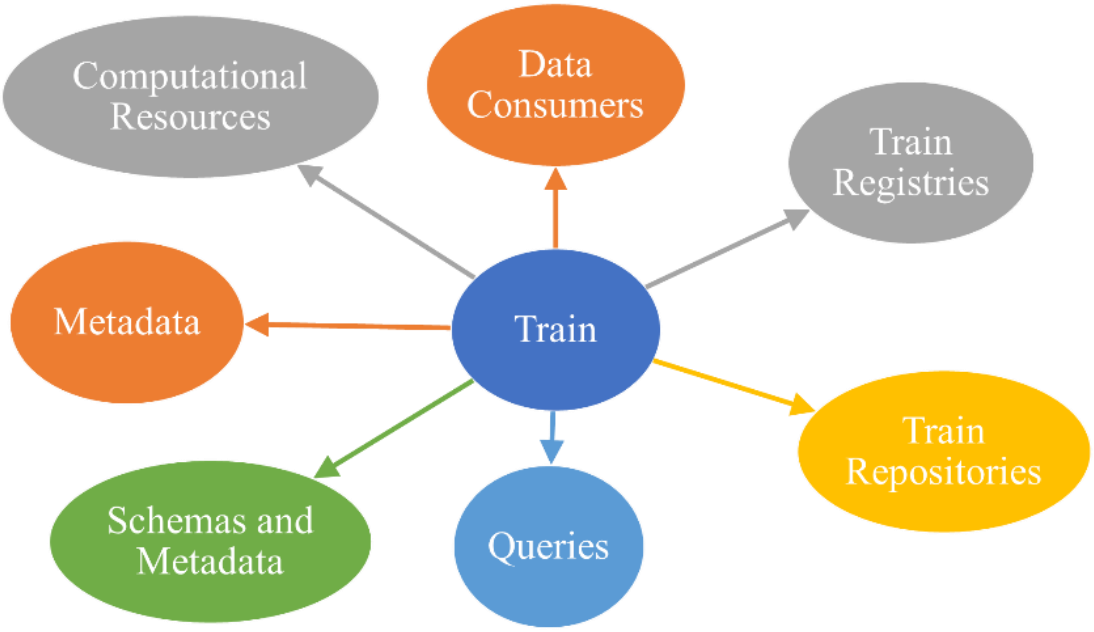


Figure 6.10 The Train

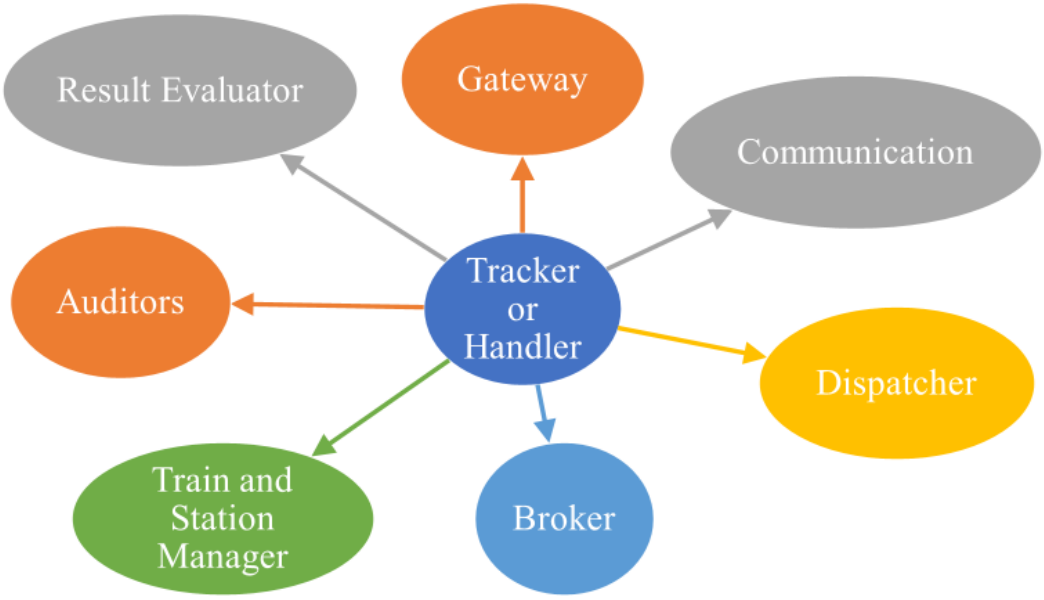


Figure 6.11 The Tracker

6.3.3 FAIR Principles for Distributed Analytics in PHT

The PHT approach promotes improving the reuse of data by sharing analytics, which can interact with the data and complete its task without giving access to the end user. Within the PHT, the FAIR principles are applied to both the Train and Station concepts, keeping in mind that the goal is enhancing the reusability of distributed data with distributed analytics.

The PHT needs to interact with data repositories, which may or may not follow FAIR principles, despite the fact that having FAIR data is highly desirable. Participating data repositories independently decide at which degree they will support FAIR data. They act as FAIR data points by implementing custom interfaces supporting the computational task that reuses data.

The PHT sets the machine readability at the core, aiming for maximal interoperability between diverse systems. Therefore, it is well aligned with FAIR principles. The components of the PHT infrastructure support FAIR principles at varying degrees as illustrated by Figure 6.12, 6.13 and Table 6.9.

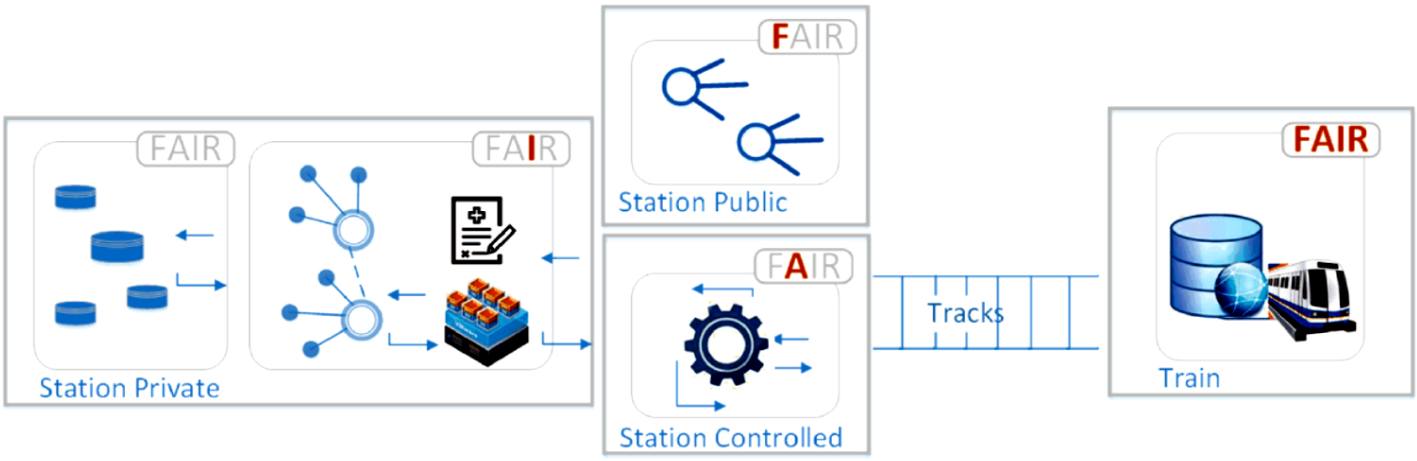


Figure 6.12: Applicability of FAIR principles to the components of the PHT Source [10]

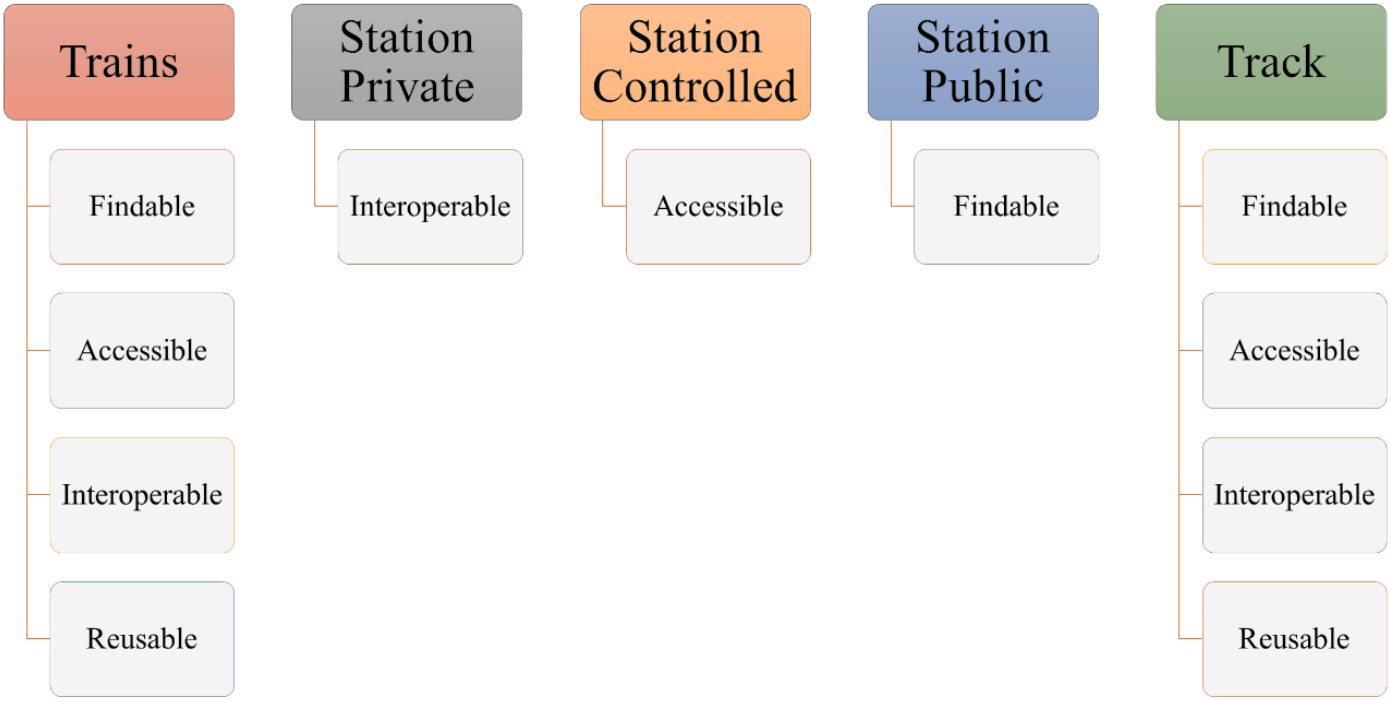


Figure 6.13: FAIR principles supported by the PHT components.

Table 6.9: FAIR principles supported by the PHT components.

PHT	Concepts Functionalities	FAIR Principles
Trains	They are data analytics tasks that are uniquely identified, richly described with metadata, registered and deposited to repositories. They are machine readable and executable digital objects.	Findable Accessible Interoperable Reusable
Station Private	Contains private data repositories and a data integration layer. Links data, stores access rights, exposes data with a standard representation by using terminologies and vocabularies.	Interoperable
Station Controlled	Executes Trains in a controlled environment. It has defined protocols to communicate with Trains and must also have authentication and authorization procedures. Log data are available after the execution of the task is complete.	Accessible
Station Public	Stations are registered in a repository with metadata. They publish both metadata about the contained data repositories and computational capabilities.	Findable
Track	Provides communication protocols and keeps track of all the communication. Supports traceability and reproducibility of the executed analytics.	FAIR

6.3.4 Summary of PHT

The PHT is a novel approach establishing a FAIR distributed data analytics infrastructure enabling the (re)use of distributed healthcare data, while data owners stay in control of their own data.

a. empowers citizens and organizations to control the use of the data that reside in their own data repositories for the benefit of the individual and society,
b. improves the usability of health data by lowering the barriers for data protection, by ensuring that the privacy and confidentiality of the data subject will be preserved,

c. ensures data sovereignty beyond data security and privacy by supporting the responsible use and builds trust between data consumers and data owners by making analytics processes repeatable, transparent and auditable,
d. applies FAIR principles to the protocols of how data analytics interacts with FAIR data points by making data analytics tasks itself FAIR and placing machine readability at its core.

The PHT provides a distributed, flexible approach to use data in a network of participants, incorporating the FAIR principles. The PHT facilitates the responsible use of sensitive and/or personal data by adopting international principles and regulations. It supports accountability by providing provenance of analytics execution and audit mechanisms.

6.3.4 Use Cases of PHT

a. The Maastro clinic has implemented a Patient Cohort Counter (PCC) "Train" as a demonstration using multiple data representations. The PCC calculates the number of matching patients and cohort statistics for a specific disease at a PHT data Station.
b. The Varian Learning Portal by Varian Medical Systems
c. The open-source software ppDLI by IKNL which are both example implementations of distributed learning PHT infrastructures in healthcare.
d. SMITH and DIFUTURE projects funded by the German Medical Informatics Initiative have developed cross consortia implementations and tested phenotyping use cases
e. The PHT approach can be applied to various other domains like agricultural sector and the courts.

Assessments – TMA	**Tutor Marked Assignment I** ***In your own words, Personal Health Train*** ***How does FAIR relate to PHT?*** ***Is FDP important to researcher?*** ***If Yes, Justify your response with facts.*** ***If No, what are the differences?***
	Explain FDP components in your own words ***Compare popular data model*** ***Explain the five-star rule of Linked data***

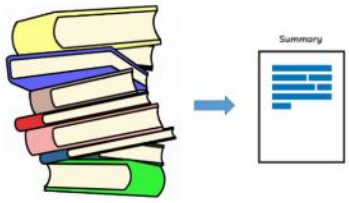
More information at this link https://youtu.be/aEVzf89xfgE, https://vimeo.com/143245835

References

1. Jansen P, van den Berg L, van Overveld P, et al. Research Data Stewardship for Healthcare Professionals. 2018 Dec 22. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science [Internet]. Cham (CH): Springer; 2019. Chapter 4. Available from: https://www.ncbi.nlm.nih.gov/books/NBK543528/ doi: 10.1007/978-3-319-99713-1_4
2. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
3. Wilkinson MD, Sansone S, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. A design framework and exemplar metrics for FAIRness. Sci Data. 2018;5:180118. https://doi. org/10.1038/sdata.2018.118.
4. Boeckhout M, Reuzel R, Zielhuis G. The donor as partner – How to involve patients and the public in the governance of biobanks and registries. Leiden: BBMRI-NL; 2014.
5. Australian National Data Service. Guide on metadata. 2016. UK Data Service. Document your data.
6. http://data4lifesciences.nl/hands/
7. How to make your data FAIR - Research Data Management Support - Universiteit Utrecht (uu.nl)
8. FAIR Data Point | FAIR Data Points can be used to describe your data sets in a FAIR way, using standard metadata and make them available through simple WWW protocols.
9. O. Beyan, A. Choudhury, J van Soest, O. Kohlbacher, L. Zimmermann, H. Stenzhorn, Md. R. Karim, M. Dumontier, S. Decker, L.O. Bonino da Silva Santos & A. Dekker. Distributed analytics on sensitive medical data: The Personal Health Train. Data Intelligence 2(2020), 96–107. doi: 10.1162/dint_a_00032
10. http://opendatahandbook.org/guide/en/what-is-open-data/
11. Luiz Olavo Bonino da Silva Santos, Mark D. Wilkinson, Arnold Kuzniar, Rajaram Kaliyaperumal, Mark Thompson, Michel Dumontier, Kees Burger. (2016) FAIR Data Points Supporting Big Data Interoperability. Enterprise Interoperability - Proceedings of the Workshops of the Eighth International Conference I-ESA. Pp 1-11

12. D.B. Deutz, M.C.H. Buss, J. S. Hansen, K. K. Hansen, K.G. Kjelmann, A.V. Larsen, E. Vlachos, K.F. Holmstrand (2020). How to FAIR: a Danish website to guide researchers on making research data more FAIR
13. https://doi.org/10.5281/zenodo.3712065
14. https://howtofair.dk/about/
15. https://dataedo.com/kb/data-glossary/what-is-metadata
16. https://falkland-cms-api.readthedocs.io/en/latest/taxonomy.html
17. Keet, C. M. (2019) The African Wildlife Ontology tutorial ontologies: requirements, design, and content. arXiv:1905.09519v1 [cs.AI]
18. https://en.wikipedia.org/wiki/Machine-readable_data
19. http://opendatahandbook.org/glossary/en/terms/machine-readable/
20. https://www.fairdatasolutions.com/fair-solutions-products#FDP
21. The DNA Bank: High-Security Bank Accounts to Protect and Share Your Genetic Identity, Johan T. den Dunnen* DOI: 10.1002/humu.22810
22. https://www.dtls.nl/fair-data/personal-health-train/

Summary of Study Unit 6



In this study unit, you have learnt that:

1. The clinical researcher is the principal data steward. He is responsible for the complete scientific process: from study design to data collection, analysis, storage, sharing and protecting the privacy of study subjects.
2. The formal responsibility for personal data lies with research institutes, which is accountable for having adequate policies, facilities, and expertise around data stewardship.
3. Decisions on data stewardship will affect how to process, analyse, preserve, and share research data in the future.
4. Explain what a researcher needs to study the design and registration; Re-use existing data, collaborate with patients, draw data management and statistical analysis plan, choose the file format; take care of Intellectual Property Rights and Data Access.

5. The clinical researcher calls for careful attention to the privacy and autonomy of people involved by getting informed consent, care and research environment and also prepare sensitive data for use.
6. The things to guide researcher when collecting data are data management infrastructure, data monitoring and validation; meta data, security, access policy and protecting research data
7. How to plan analysis and archiving data.
8. The Personal Health Train (PHT) proposes an alternative approach to existing data sharing and licensing agreement to which encompasses both technological and social aspects of sensitive data reuse.
9. The PHT does not require the transfer of data from the holding entity. Rather than moving the data to the requester, it moves the analytics tasks to the data repositories and executes the tasks in a secure environment.
10. The core components of PHT are Station, Train and Handler
11. The PHT approach promotes improving the reuse of data by sharing analytics, which can interact with the data and complete its task without giving access to the end user.
12. Within the PHT, the FAIR principles are applied to both the Train and Station concepts, keeping in mind that the goal is enhancing the reusability of distributed data with distributed analytics.
13. On applicability of FAIR principles to the components of the PHT, we have Station Private, Station Controlled; Station Public; Trains and PHT Track
14. Listed some use cases of PHT

Self-Review Questions for Study Unit 4

Now that you have completed this study unit, you can assess how well you have achieved its Learning Outcomes by answering these questions.

1. Describe some of the roles of PHT
2. Explain how FAIR supports PHT
3. Briefly describe the roles of each component of PHT

4. Compare Open and FAIR Data
5. Describe briefly the components of FDP
6. List some benefits of FDP to Researchers

Self-Review Answers (SRA) to Self-Review Questions of Study Unit 4

1. PHT
   a. empowers citizens and organizations to control the use of the data that reside in their own data repositories for the benefit of the individual and society,
   b. improves the usability of health data by lowering the barriers for data protection, by ensuring that the privacy and confidentiality of the data subject will be preserved,
   c. ensures data sovereignty beyond data security and privacy by supporting the responsible use and builds trust between data consumers and data owners by making analytics processes repeatable, transparent and auditable,
   d. applies FAIR principles to the protocols of how data analytics interacts with FAIR data points by making data analytics tasks itself FAIR and placing machine readability at its core.
2. FAIR principles supported by the PHT components

PHT	FAIR principles
Train	FAIR
Station Private	Interoperable
Station Controlled	Accessible
Station Public	Findable
Track	FAIR

3. Station provides curated, confidential data and acts as FAIR data points. Stations expose data in a discoverable format, define an interface to execute queries, provide computational resources and execute analytic tasks in a secure environment. A Train carries different components; namely, metadata that stores the Train's unique digital persistent identifier, study

description, the query used in data extraction, analytics for data utilization and aggregation for result integration. The Track or Handler It manages Train and Station states and logs the transaction information for future auditing.

4.

Open data	FAIR Data
It is available to everyone to access, use, and share, without licenses, copyright, or patents	It uses the term "Accessible" to mean accessible by appropriate people, at an appropriate time, in an appropriate way. This means that data can be FAIR when it is private, when it is accessible by a defined group of people, or when it is accessible by everyone (open data). It depends completely on the purpose of the data, where the data currently is in its lifecycle, and the end-usage of the data
An example is an undocumented data dump in an uncurated repository, such as OSF, which is neither findable, nor reuseable, nor interoperable)	An example is a data set that is findable, reuseable, etc., but only accessible within a closed research group

5.

Components	Description
Metadata Provider	This is responsible for giving access to the metadata.
FAIR Accessor	The FAIR Accessor component provides access to the actual data content of the dataset.
Metrics Gatherer	The Metrics Gatherer component monitors various aspects the FDP usage.
Security Enforcer	The Security Enforcer component should act as a gatekeeper, protecting the access to the (meta)data from requests that do not comply with the given licenses.

6. Benefits of FDP for research(ers)

a. Making research data more FAIR will provide a range of benefits to researchers, research communities, research infrastructure facilities and research organisations alike, including:
b. Achieving maximum impact from research.
c. Increasing the visibility and citations of research.
d. Improving the reproducibility and reliability of research.
e. Attracting new partnerships with researchers, business, policy and broader communities.
f. Enabling new research questions to be answered