Dissertation

In

Elucidata Corporation

(March 7th to August 6th, 2022)

Implementing FAIR data principles in biomolecular data through Polly

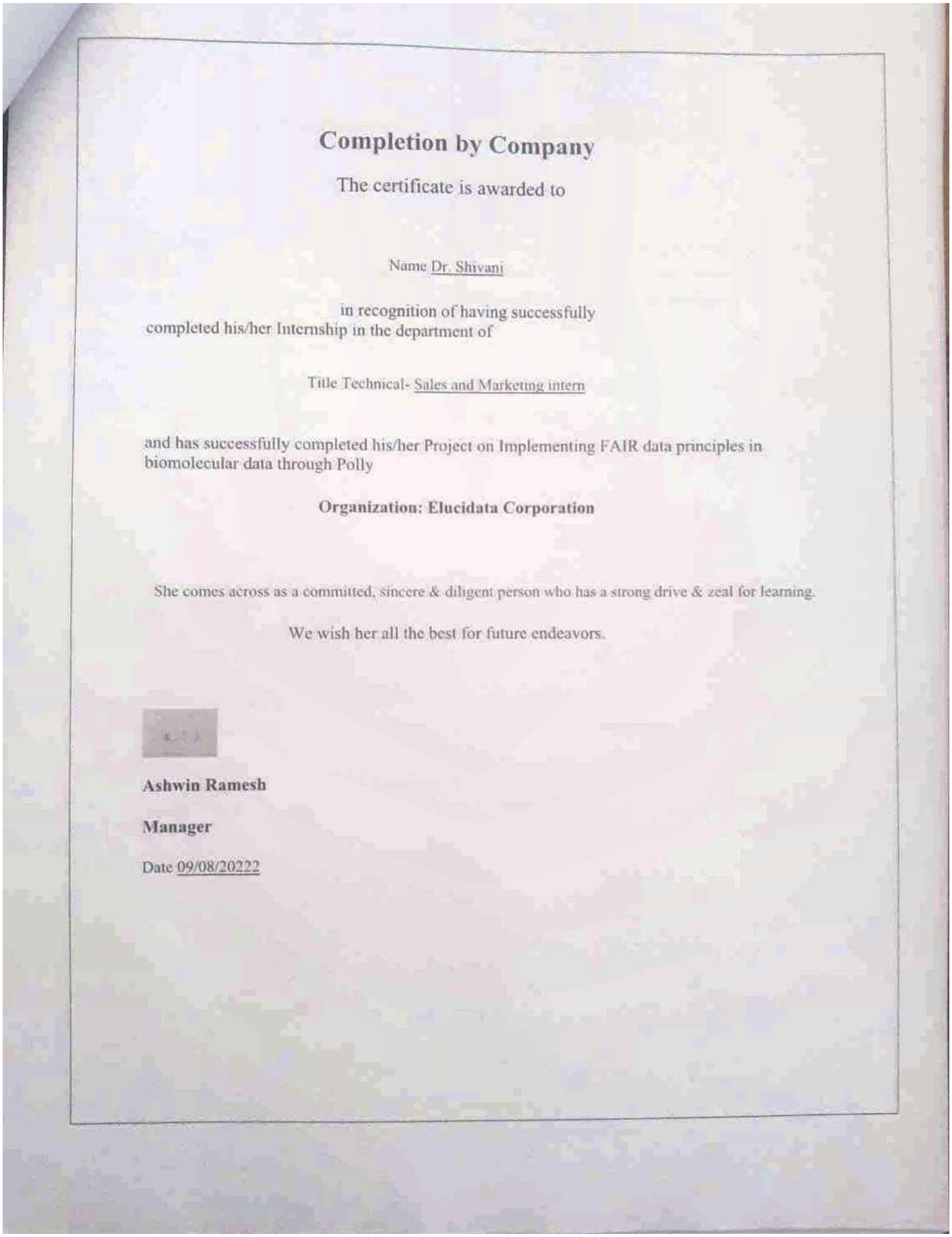
Submitted by

Dr. Shivani

(PG/20/074)

Post Graduate Diploma in Hospital and Health Management

IIHMR DELHI International Institute of Health Management Research, New Delhi



Completion by Company

The certificate is awarded to

Name Dr. Shivani

in recognition of having successfully completed his/her Internship in the department of

Title Technical- Sales and Marketing intern

and has successfully completed his/her Project on Implementing FAIR data principles in biomolecular data through Polly

Organization: Elucidata Corporation

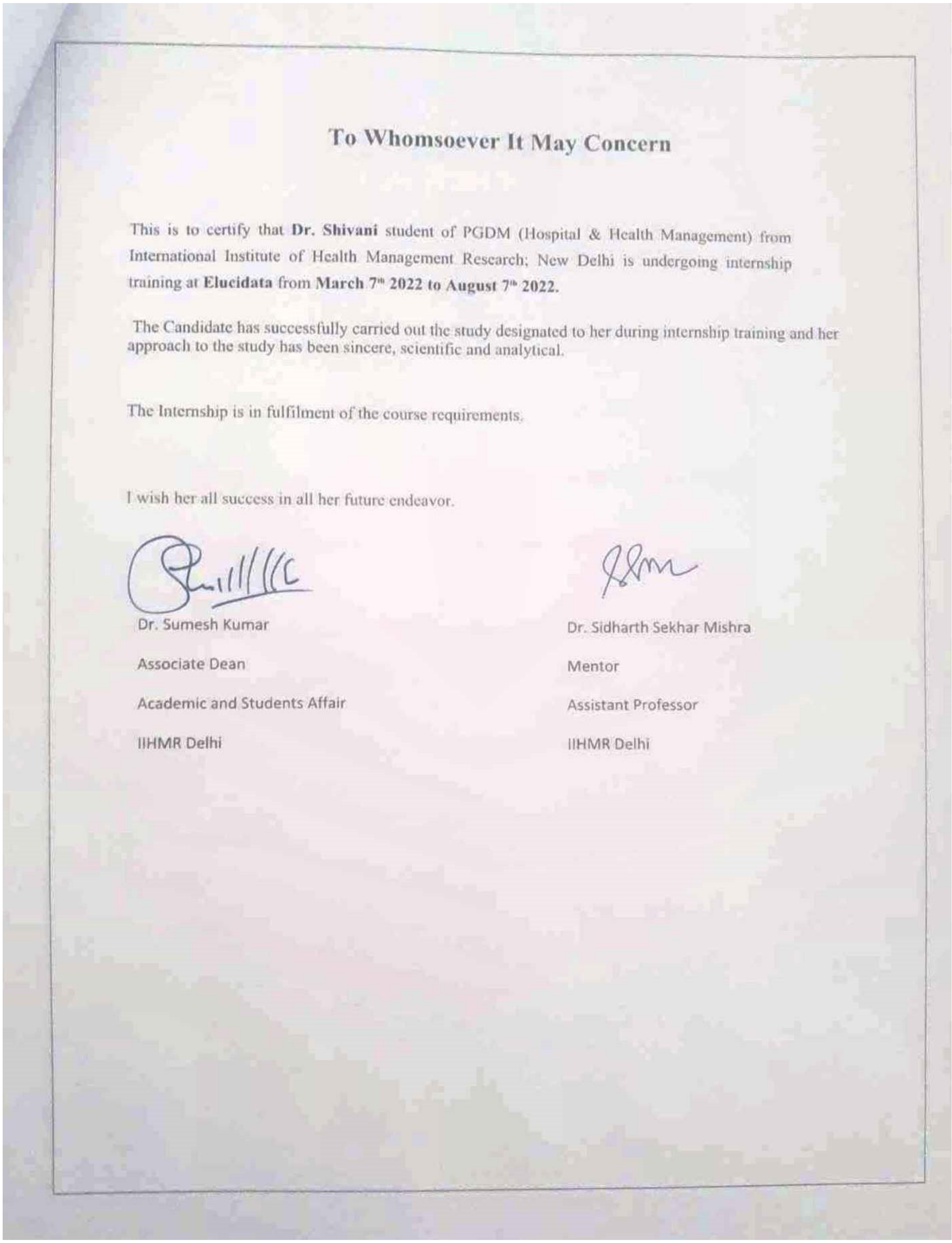
She comes across as a committed, sincere & diligent person who has a strong drive & zeal for learning.

We wish her all the best for future endeavors.

Ashwin Ramesh

Manager

Date 09/08/20222



To Whomsoever It May Concern

This is to certify that **Dr. Shivani** student of PGDM (Hospital & Health Management) from International Institute of Health Management Research; New Delhi is undergoing internship training at **Elucidata** from **March 7th 2022 to August 7th 2022.**

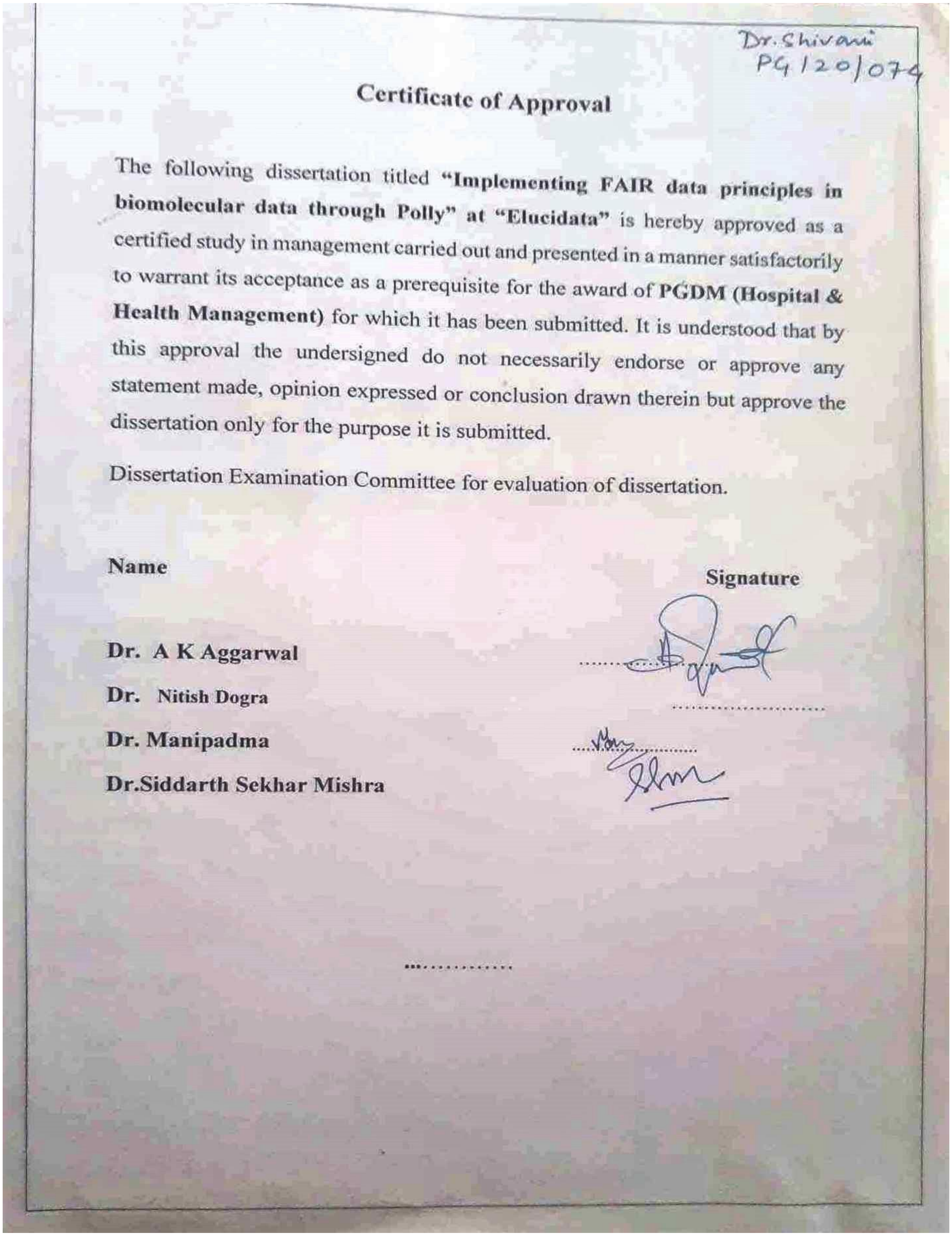
The Candidate has successfully carried out the study designated to her during internship training and her approach to the study has been sincere, scientific and analytical.

The Internship is in fulfilment of the course requirements.

I wish her all success in all her future endeavor.

Dr. Sumesh Kumar
Associate Dean
Academic and Students Affair
IIHMR Delhi

Dr. Sidharth Sekhar Mishra
Mentor
Assistant Professor
IIHMR Delhi



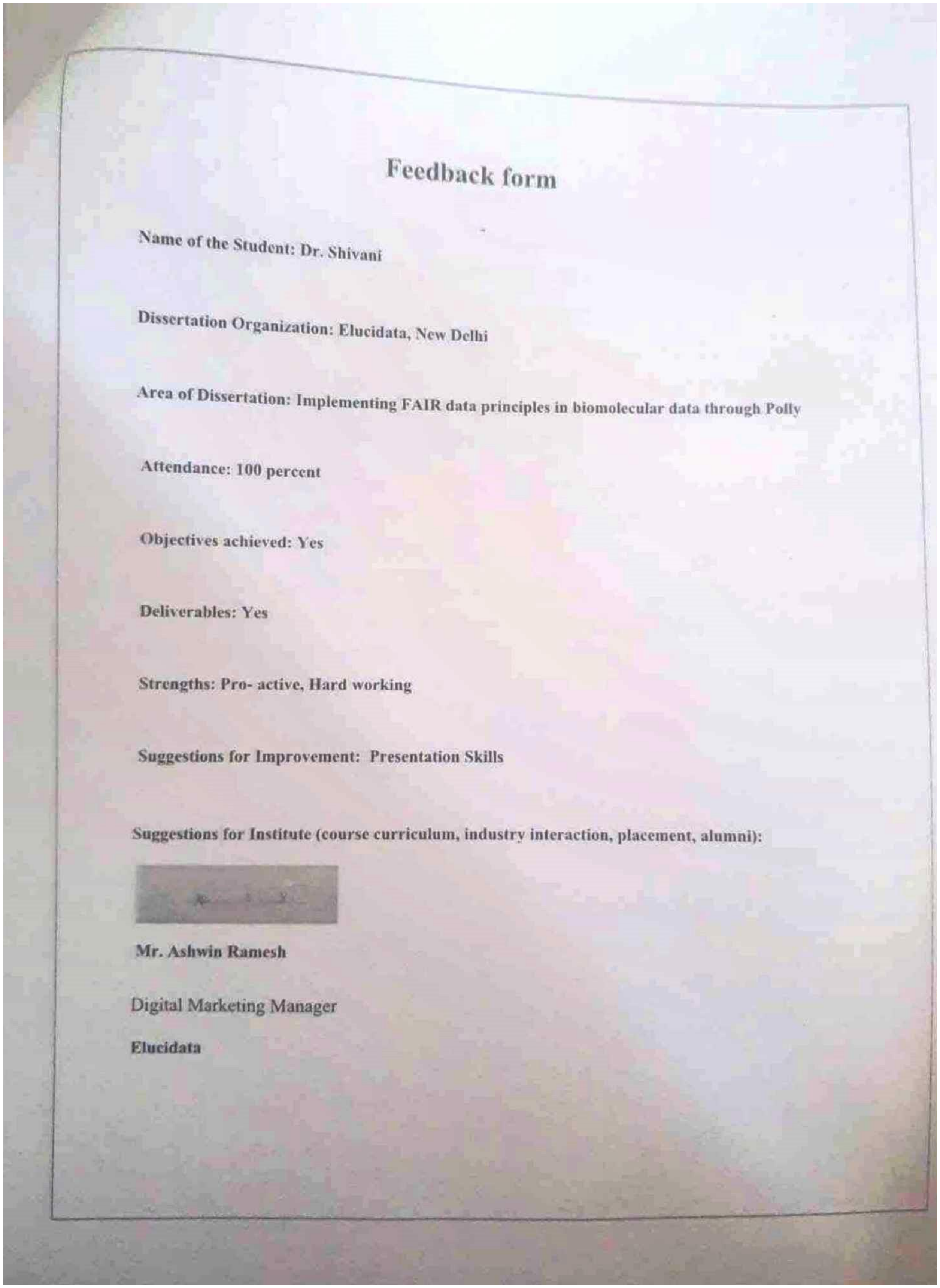
Dr. Shivani
PG/20/074

Certificate of Approval

The following dissertation titled **"Implementing FAIR data principles in biomolecular data through Polly" at "Elucidata"** is hereby approved as a certified study in management carried out and presented in a manner satisfactorily to warrant its acceptance as a prerequisite for the award of **PGDM (Hospital & Health Management)** for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the dissertation only for the purpose it is submitted.

Dissertation Examination Committee for evaluation of dissertation.

Name	**Signature**
Dr. A K Aggarwal	
Dr. Nitish Dogra	
Dr. Manipadma	
Dr.Siddarth Sekhar Mishra	



Feedback form

Name of the Student: Dr. Shivani

Dissertation Organization: Elucidata, New Delhi

Area of Dissertation: Implementing FAIR data principles in biomolecular data through Polly

Attendance: 100 percent

Objectives achieved: Yes

Deliverables: Yes

Strengths: Pro- active, Hard working

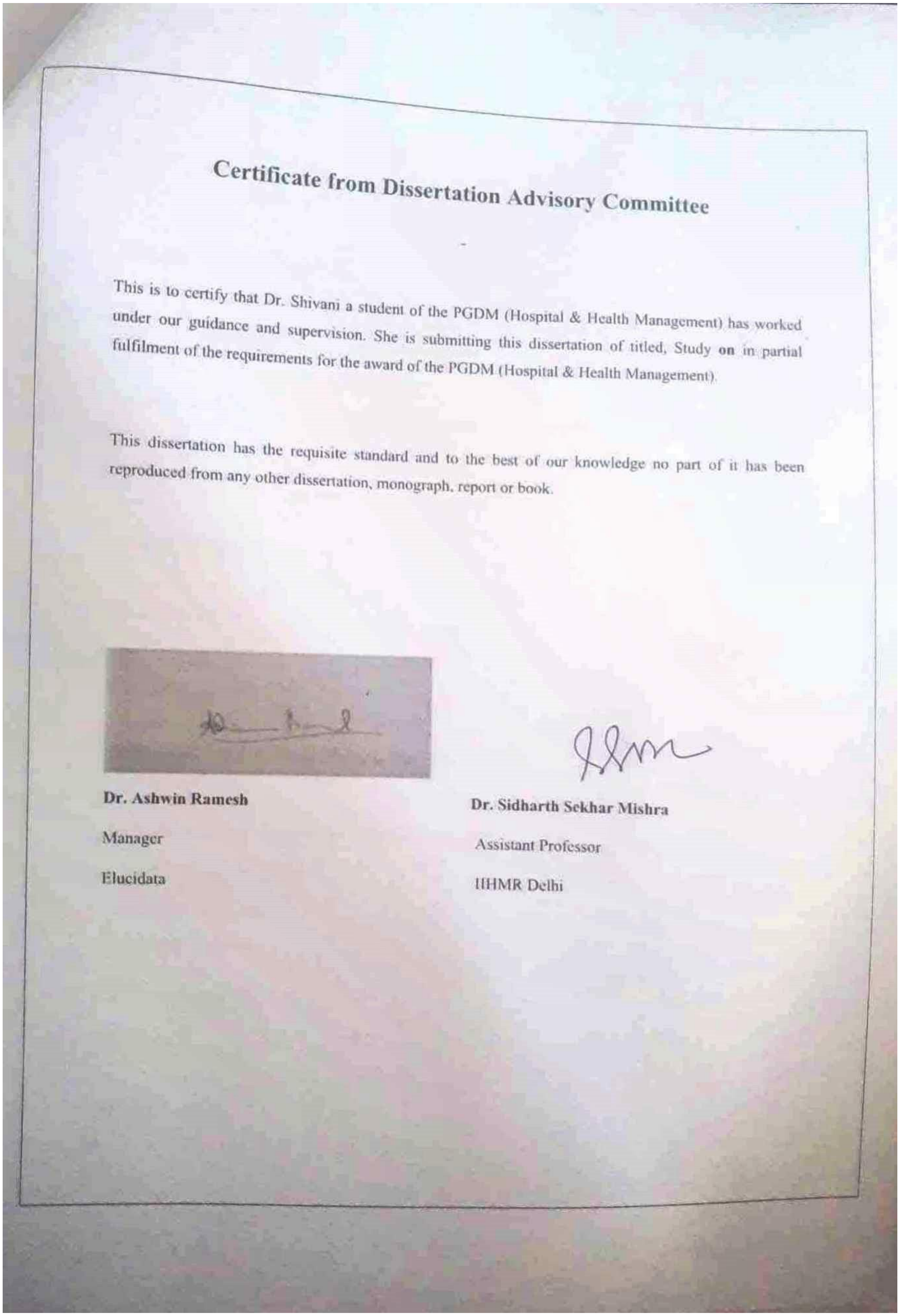
Suggestions for Improvement: Presentation Skills

Suggestions for Institute (course curriculum, industry interaction, placement, alumni):

Mr. Ashwin Ramesh

Digital Marketing Manager

Elucidata



Certificate from Dissertation Advisory Committee

This is to certify that Dr. Shivani a student of the PGDM (Hospital & Health Management) has worked under our guidance and supervision. She is submitting this dissertation of titled, Study **on** in partial fulfilment of the requirements for the award of the PGDM (Hospital & Health Management).

This dissertation has the requisite standard and to the best of our knowledge no part of it has been reproduced from any other dissertation, monograph, report or book.

Dr. Ashwin Ramesh

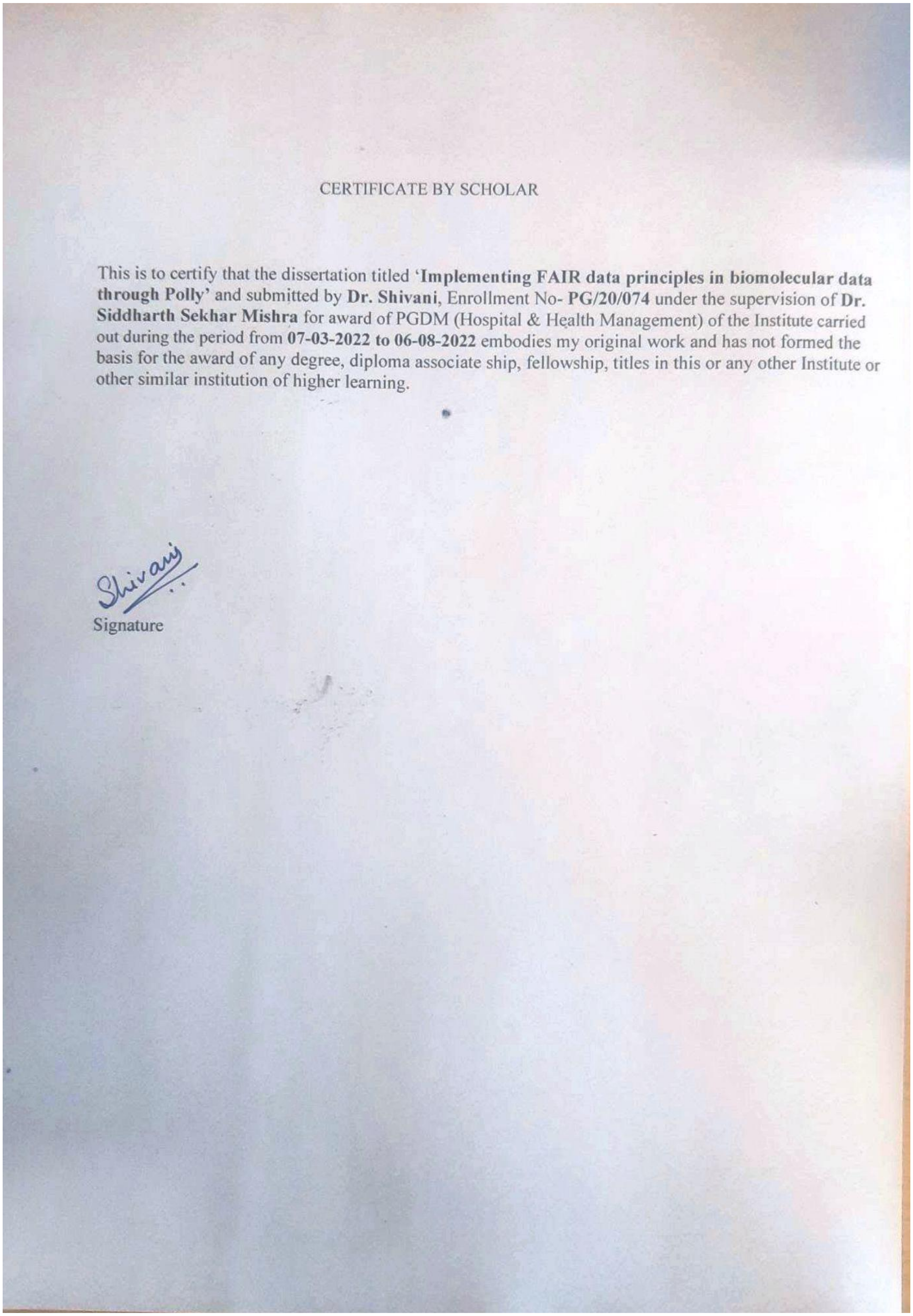
Manager

Elucidata

Dr. Sidharth Sekhar Mishra

Assistant Professor

IIHMR Delhi

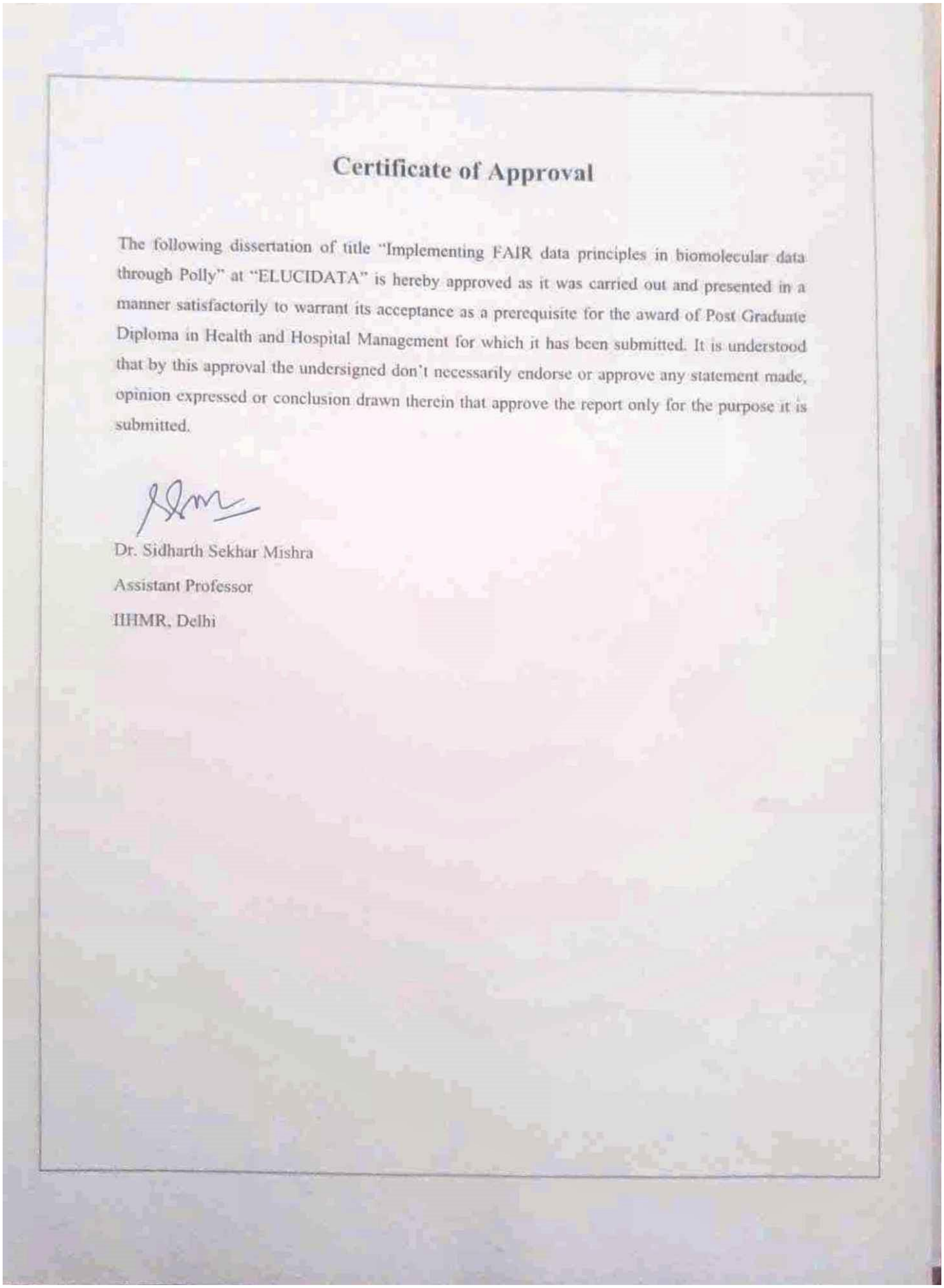


CERTIFICATE BY SCHOLAR

This is to certify that the dissertation titled '**Implementing FAIR data principles in biomolecular data through Polly'** and submitted by **Dr. Shivani**, Enrollment No- **PG/20/074** under the supervision of **Dr. Siddharth Sekhar Mishra** for award of PGDM (Hospital & Health Management) of the Institute carried out during the period from **07-03-2022 to 06-08-2022** embodies my original work and has not formed the basis for the award of any degree, diploma associate ship, fellowship, titles in this or any other Institute or other similar institution of higher learning.

Shivani

Signature



Certificate of Approval

The following dissertation of title "Implementing FAIR data principles in biomolecular data through Polly" at "ELUCIDATA" is hereby approved as it was carried out and presented in a manner satisfactorily to warrant its acceptance as a prerequisite for the award of Post Graduate Diploma in Health and Hospital Management for which it has been submitted. It is understood that by this approval the undersigned don't necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein that approve the report only for the purpose it is submitted.

Dr. Sidharth Sekhar Mishra
Assistant Professor
IIHMR, Delhi

Acknowledgments

The internship opportunity I had with Elucidata was an excellent chance for learning and professional development. Therefore, I consider myself a very lucky individual as I was provided with an opportunity to be a part of it.

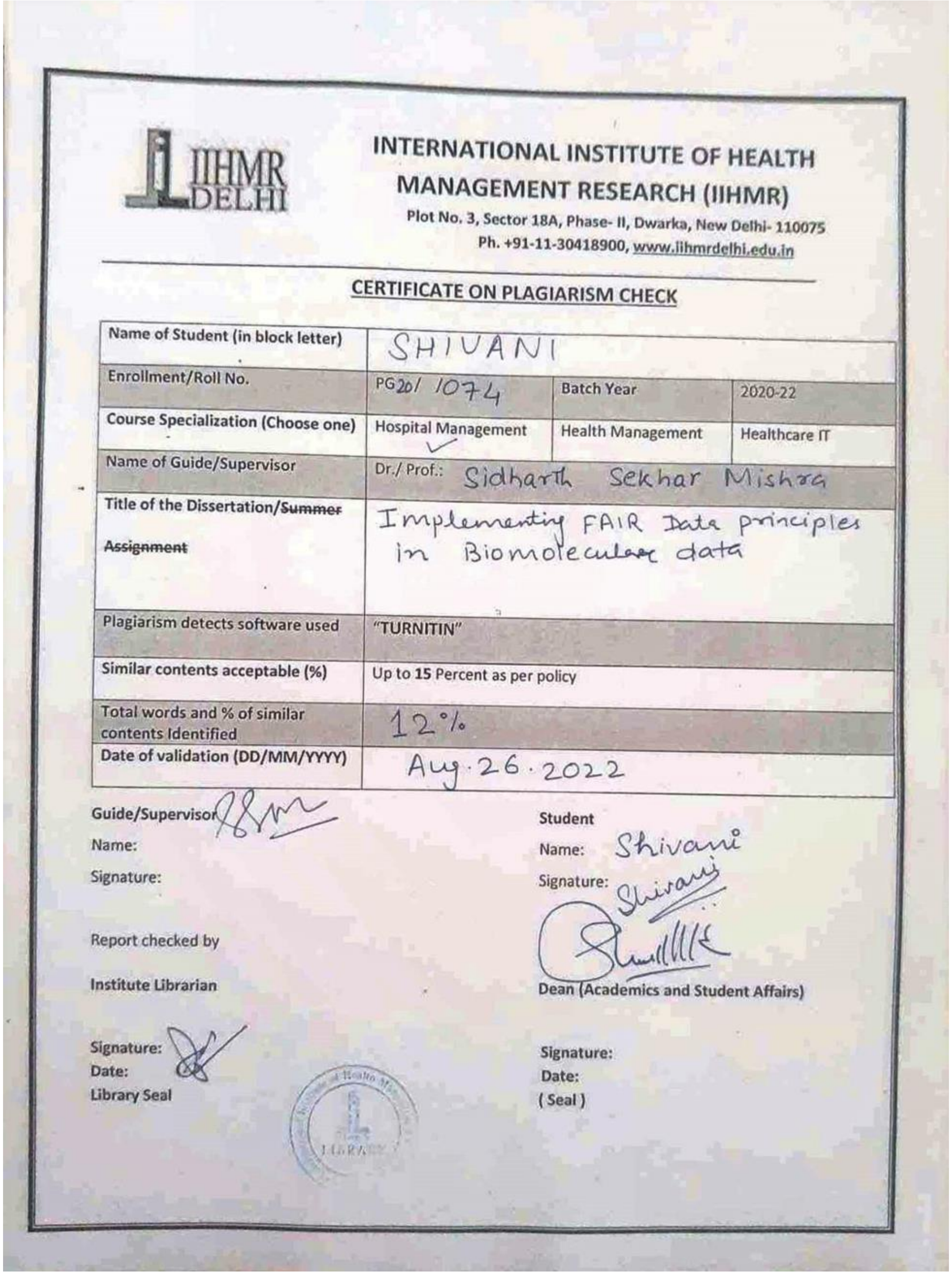
I am using this opportunity to express my deepest gratitude and special thanks to Mr. Ashwin Ramesh (Digital Marketing Manager- Elucidata), who, despite being extraordinarily busy with his duties, took time out to hear, guide, and keep me on the correct path and allowing me to carry out my project at their esteemed organization and extending during the training.

It is my radiant sentiment to record my best regards, most profound sense of gratitude to my mentor Dr. Sidharth Shekhar Mishra (Assistant professor, IIHMR Delhi), for his valuable guidance and co-operation in my endeavor.

I perceive this opportunity as a significant milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement to attain my desired career objectives.

Shivani
9.8.2022

Dr. Shivani



INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT RESEARCH (IIHMR)

Plot No. 3, Sector 18A, Phase- II, Dwarka, New Delhi- 110075
Ph. +91-11-30418900, www.iihmrdelhi.edu.in

CERTIFICATE ON PLAGIARISM CHECK

Name of Student (in block letter)	SHIVANI		
Enrollment/Roll No.	PG20/ 1074	Batch Year	2020-22
Course Specialization (Choose one)	[✓ Hospital Management]	Health Management	Healthcare IT
Name of Guide/Supervisor	Dr./ Prof.: Sidharth Sekhar Mishra		
Title of the Dissertation/~~Summer Assignment~~	Implementing FAIR Data principles in Biomolecular data		
Plagiarism detects software used	"TURNITIN"		
Similar contents acceptable (%)	Up to **15** Percent as per policy		
Total words and % of similar contents Identified	12%		
Date of validation (DD/MM/YYYY)	Aug.26.2022		

Guide/Supervisor
Name:
Signature:

Student
Name: Shivani
Signature: Shivani

Report checked by

Institute Librarian

Dean (Academics and Student Affairs)

Signature:
Date:
Library Seal

Signature:
Date:
(Seal)

Table of contents

List of the tables

List Of the figures

Abbreviations

FAIR	Findable, Accessible, Interoperable and reusable
BBMRI-ERIC	Biobanking and BioMolecular resources Research Infrastructure (**BBMRI**)- European Research Infrastructure Consortium (**ERIC**)
GA4GH	Global Alliance for Genomics and Health
BRCA	BReast CAncer gene
DepMap	Dependency Map
LINCS	Library of Integrated Network-based Cellular Signatures
GEO	Gene Expression Omnibus
TCGA	The Cancer Genome Atlas
ELISA	Enzyme-Linked Immunoassay
GCT	Gene Cluster Text

Implementing FAIR data principles in biomolecular data through Polly

Introduction:

The relevance of well-managed research data from academic and industrial sectors stretches well beyond the original research goal. Despite the fact that the evident benefits of data generated, the life sciences data is neither easily discoverable nor accessible or interoperable, rendering it non-reusable.

Science is distinguished for its transparency, openness, and reproducibility. As a result, these traits are supposed to be ubiquitous in everyday practice, although data demonstrates that this is not the case.

Traditional research methodology has primarily focused on generating data for the publication of articles, which are soon forgotten once they have served their purpose. Additionally, no globally known standards or guidelines allow for open research practices. It hence necessitates the development of public and private capacity for high-quality data management so that fair principles are entrenched in the data generated.

Biomedical data is not fair, and making this data fair is a challenge.

Aim:

The study's main aim is to understand the FAIR data principles and explore the role of the Polly platform in delivering ML ready biomolecular data.

Objectives:

- To outline existing scientific knowledge available on FAIR data in biomedical science
- To highlight the process of delivering ML ready data through Polly

Research Methodology

Research design: Exploratory study

Data type: Secondary data

Data collection: Literature review

Data source: Published articles, Data from Organization

Key words: FAIR Data, FAIRIfication, Metadata, Biomedical data

Need for FAIR data:

Every day, academic labs and institutions worldwide produce humongous amounts of multi-omics data through experimental works and deposit it in public repositories.

This data holds potency for reuse and discovery but is dispersed across multiple disparate sources and lacks standardization.

Additionally, no single type of data, whether metabolomic, proteomic or genomic, will be sufficient to capture the complexity of a biological phenomenon.

Adopting an integrated approach could significantly aid the ability to gain a holistic and more accurate understanding of physiology and disease pathology at the molecular level [1] [2].

Un- FAIR data practices impede Data usability

1. The lack of metadata annotations prevents data reuse and affects the findability of relevant data
2. Inconsistent data schema and processing reduce interoperability and reusability
3. Lack of infrastructure to process and store data at scale
4. Data in Silos hampers accessibility

Realizing the value of data assets in Pharmaceutical industry

The pharmaceutical industry's research and development of novel medications can be expedited by the efficient utilization of massive data leveraging cutting-edge technology as artificial intelligence (AI), including machine learning. [16-18]

Implementing efficient data management strategies that make use of the FAIR guiding principles and the related framework and standards for measuring FAIRness is vital for the pharmaceutical sector. [19]

Implementing FAIR has the potential to significantly improve the adequacy and efficiency of drug discovery, therapeutic development in the biopharmaceutical sector. Making data assets within biopharma FAIR and public data more discoverable and accessible for both machines and people will enhance the value of those resources.[20]

Fair data is essential for driving innovation in the R&D of pharmaceutical industries.

If the pharmaceutical industry wants to seize the power of Artificial intelligence and machine learning in research and development, it must incline itself to fair data. The constraint in achieving innovation in drug discovery can be the accessibility of the data on which AI relies because those data must be machine-accessible and machine-readable, which necessitates the data to be FAIR, rather than compute power or the availability of AI tools.

A range of advantages come from incorporating the FAIR principles into a company's data management policy, including the potential for process automation due to machine readability which will permit reuse. It will be easier and more efficient to flow data from acquisition to incorporation to semantic alignment to analytics to provide answers. The amount of time spent preparing data for analysis through data filtering will be reduced. Responses to scientific questions will come more quickly. As a result, the output will improve, time to value will be dropped drastically, and pharmaceutical R&D can move faster.

The benefits for the biopharma industry will be substantial, and they include:

- accelerating innovation due to the accessibility of FAIR data for use
- cutting the time from drug discovery to market value by expediting clinical studies
- constructing more-personalized medicines by leveraging FAIR real-world data to complement appropriate treatment to pertinent patient cohorts; and
- empowering data sharing and collaborative partnerships across industries.

Expected Business Impact

It will be necessary to quantify the predicted business impact of implementing FAIR in terms of ROI. In the short term, this can be assessed in terms of time and money savings to enable:

(i)greater findability for existing data.

(ii) rapid access to the data

(iii) easier access to harmonized and high-quality data for data analysis.

After the deployment of FAIR, the RoI parameters will be based on lower costs and quicker processing times to identify and develop better therapeutic treatments, leading to a perceptible rise in the productivity of the R&D pipeline.

Global data FAIRification programmes have been started by some businesses. For instance, one of the organizations developed a FAIR platform for 3000 users across three major sites. Upon running the new FAIR platform for two months, the firm gathered usage data based on user clickthrough rates. In 60 days, 900 000 pages on this FAIR platform were visited. An estimate of approximately 5 million page visits was provided for the year. By making improved search results available with direct access to the target repository, each of these FAIR-enhanced views saved approximately five seconds, which led to a computation of around 3.5 full-time employees' worth of time saved annually.

Another instance involved a pharmaceutical organization, where a production plant was taken out of action for months. When the necessary specialized data were finally located, it was revealed that the problem had already happened three times and that using FAIR's "findability" would have enabled a much more timely resolution.[20]

Cost-Benefit Analysis of FAIR data in research [21]:

The lack of FAIR data in the research industry affects research activities and opportunities for further research and innovation.

Indicators to gauge the impact of non- FAIR data on research:

- **Time spent**

  According to the percentage of time spent on research tasks, the overall time inefficiencies caused by the lack of FAIR data have been calculated for academic researchers to be 3.12 per cent. For non-academic researchers, the time inefficiency estimated is 4.47% of their total time that could be utilized for research.

  Impact in terms of cost:

  Time lost as a result of non-FAIR research costs €4.5 billion annually.

  The total cost of time lost by Non-Academics is estimated to be €3.2 billion.

  The total cost of time lost by Academics is €1.2 billion

- **Cost Storage**

  The FAIR data would save storage costs for publishers and data repositories by obviating the need for superfluous versions. The cost of keeping data is proportionate to the amount of data saved. Implementing fair data principles can reduce data redundancy by reducing data storage across various repositories by 20% (data backup is considered a security measure and not considered data redundancy).

  Impact in terms of cost

  Redundant storage owing to non-fair data consumes €5.3 billion a year.

- **License cost:**

  If the report's or the data's metadata contains a license, it may be one of the following types:

  An open licence;

  A restricted license that charges a fee for data reuse;

A localized human-readable license.

It is evident that about 45% of data is in open access today, and there is a scope of increasing it to 70 per cent, which will enhance the reuse of data by researchers and make use of data that remain in silos. The remaining data must remain accessible with a license for privacy and security purpose.

Impact in terms of Cost:

Three hundred sixty million euros are spent annually on licence fees due to the lack of open access.

- **Research Retraction:**

  There are multiple reasons for research retraction- Errors, non-reproducibility, multiple submissions, plagiarism etc.

  The amount of retracted articles corresponds to nearly 0.05 per cent, and implementation of FAIR principles can bring this down by 50%.

  Impact in terms of cost:

  Retraction of Research due to non-FAIR research costs more than €4.million annually.

- **Double Funding:**

  Anti-plagiarism technologies that automatically match research with data and current publications are largely used to prevent double financing. The effectiveness of this strategy depends on the availability of the plagiarised research in a machine-readable format, which would be made possible by following the FAIR principles.

  It is believed that FAIR could prevent at least 80% of duplicate financing because some people would undoubtedly come up with new, inventive ways to commit plagiarism.

  Impact in terms of cost:

  The annual cost of financing that is doubled due to non-FAIR research is more than €25 million.

- **Interdisciplinarity:**

  In contrast to the value of research that would be conducted in the absence of the FAIR principles, an interdisciplinary approach refers to the added value of new research that combines various fields of study.

  Impact in terms of innovation:

  Researchers may acquire fresh insights and share information more easily if they have access to diverse data from other disciplines attributable to FAIR.

- **Potential economic expansion:**

  FAIR has a favourable effect on development, which fuels the growth of jobs and a higher GDP by maximizing the value of research and furthering science.

  Impact of FAIR data

  By 2020, it was predicted that the anticipated economic benefits of open data would total between €11.7 billion and €22.1 billion in Europe.

FAIR data stewardship

The FAIR guiding Principles:

FAIR stands for Fair, Accessible, Interoperable and reusable.

F: Findable

Data are described with rich metadata, and metadata explicitly includes the identifier of the data it describes. The metadata is registered or indexed in searchable resources and is assigned a globally unique and persistent identifier.

E.g., BBMRI-ERIC Directory

A: Accessible

(meta)data are retrievable by their identifier using a standardized communication protocol. The protocols are open, free and universally implementable. The protocol must allow for authorization and authentication processes whenever required.

E.g., European genome–phenome archive.

I: Interoperable

(meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation. (meta)data use vocabularies that follow FAIR principles.

E.g., GA4GH Genomic Data Toolkit

R: Reusable

metadata is thoroughly specified with a variety of precise and important attributes. Metadata is provided with a comprehensible data usage policy. Accurate origin is coupled with metadata. Metadata adheres to community norms pertinent to the domain.

E.g., BRCA exchange [3] [4].

Ethical Values of FAIR:

Values underlying FAIR principles:

Findability:

- Better and more inclusive research: Findable data facilitates data access and reuse by removing obstacles to data sharing amongst research groups.
- Transparency: For data to be found, the data owner must offer information about the source, processing, and sharing of their FAIR data (metadata).
- Economical: FAIR data lower economic and energy expenses related to data production, helping preserve a decent environment for present and future generations by facilitating research replication for validation and minimizing unnecessary duplication.
- The legitimacy of a number of tasks: by making data more accessible, the job of quality data generation becomes more acknowledged, whereas if data are not findable, it might be overlooked as an accomplishment or a study result.

Accessibility

- Accountability: Data providers can leverage accessible data to establish a fair and unambiguous redistribution of duties along the data reuse pipeline. To make these standards relevant, data providers must provide practical approaches.
- Trust: data providers can only make their data accessible on the background of confidentiality between them and stakeholders, including patients, data users and the medical community.
- The benefit to the community: Data providers might share their data with the scientific community and the general public in order to accomplish a common benefit via knowledge.
- Value addition to the health: Data can be made accessible to the research community in order to aid in the development of safer technology and to enhance patient outcomes (public health). The goal is to redesign healthcare delivery systems to maximize the benefits of care for patients.

Interoperability:

- Interoperability in the data can help to ensure that everyone has equal access to information and benefits from data. This can also assist the distribution of the benefits and opportunities, particularly with regard to deprived scientific or societal groups (i.e. due to economic, technical, or geographic disadvantages).
- Reproducibility: Interoperability of data is necessary for essential verification and replication of results. The use of available technological and organizational solutions is made easier, which saves time and resources.
- Novelty: the interoperability of data allows for a larger sample size which in turn paves the way for newer questions
- Quality: Interoperability of data helps to improve the quality of findings.

Reusable:

- Reciprocity: data providers can make their data reusable and, as a result, gain acknowledgement from the research community and/or the general public for their work that extends beyond their anticipations.
- Benefit-sharing: reusing excellent quality data enhances the advantage of the efforts to create these data in terms of research output, and several publications will result from the (re) usage of the same data.
- Non - maleficence: Because data reuse requires respect for privacy and data protection, the "non-nocere" concept lies at the core of the system. At the same time, FAIR data must preserve people's privacy and data while also supporting a societal good associated with research utilization for the benefit of society.
- Long-term perspective and generosity: data reuse provides the initial labour done to create the data a temporal ordering that goes toward eternity.
- Freedom of research is strengthened by enabling data to be reused for different research projects, which would otherwise be impossible.
- Human dignity and autonomy: allowing data to be reused can help make sure that everyone's dignity is protected. Respect for human freedom is implicit in the autonomy principle. In that sense, it means that people must be allowed to choose whether or not to disclose their data. Individuals must also be informed of the conditions of the data collecting and sharing procedure, the aims of data sharing, and the fact that they can

withdraw their consent at any moment. At the same time, being aware of numerous research uses might create the impression that their engagement in research is more important to society [5].

Metadata-data about the data:

A key element in FAIR principles is that metadata and its standards should be enunciated and made available to the scientific community and other stakeholders to the greatest extent possible. Metadata in a comprehensible language refers to systematic descriptions and attributes of datasets pertinent to interpreting data in an intelligible manner.

Shades of metadata [6]:

1. Descriptive metadata: This is the simplest form of metadata- a mall string describing a thing for, e.g., title and description.
2. Structural metadata: Structural metadata is concerned with giving descriptors which allow users to understand how data is organized. For instance, this sort of metadata could provide information about a table's architecture, relationships between items, and types.
3. Administrative metadata: this metadata covers things like record identifier
    1. Provenance metadata: Processes such as origination, alteration, and conversion should be detailed, along with the dates and individuals who carried them out. Versioning data is also included in this type of metadata data. This information was recognized as "Audit and Trail" information because it helps users to understand how data was created and so gives trust tokens and documentation that may be used to validate data's authenticity, dependability, and credibility.
    2. Legal metadata: This subclass of metadata was focused on supplying tags that allowed information about the data's restriction of use, copyright information, and intellectual coverage to be provided.
4. Quality metadata: This might be a relatively new component which has been introduced to the metadata collection. It provides distinctive elements for establishing indicators that disclose information about the quality of a data. That can include quantitative measures like a variable's variance or standard error, as well as more qualitative characteristics like a discrete ranking.

FAIRification:

For a variety of reasons, including concerns regarding scientific integrity, reproducibility, and transparency, as well as new needs and opportunities for large-scale data analysis and reanalysis, calls for expediting broader access and reuse of scientific data have rapidly picked up steam throughout health and biomedical research [7] [8] [9].

"FAIRification" is a term coined to describe the process of making data discovery and reusing FAIR. The FAIR guiding principles for research data management, which stand for Findability, Accessibility, Interoperability, and Reusability, might be a step in the right direction.

The principles were developed in 2014 as a framework of guiding principles and practises for stewardship of scientific data in the biomedical sciences. The concepts have acquired a lot of momentum in research and research policy since then. These are expected to become a pillar of research policy and research data management plan standards [4].

Machine actionability:

Enhancing automation is a necessary precursor for research. For reliable processing of sensor data, automating the data retrieval from data repositories through APIs is crucial. Each one of the FAIR principles requires the use of computers and computer-assisted data management and analysis. Machine-actionability extends to genome-level variant calling along with aggregate-level data and biobank databases at all levels of data aggregation [10].

Controlled access:

The FAIR principles urge that the terms and conditions under which data is shared or made accessible be clear, well-defined, and readily available. The FAIR principles essentially attempt to make foundation requirements for data sharing clearly evident, such as criteria for acquiring and providing data access, privacy, and publishing and usage barriers. [ref]

These indicate FAIR principles are compatible with models of controlled data access and release. By offering a middle ground to which more stakeholders may comply, the FAIR principles provide a way out of the dilemmas of coupling open research with the interests and values of confidentiality and intellectual property. Rather than advocating for the open and free access in general, the goal is to find appropriate and effective ways to regulate access while supporting legitimate research for all data. This is in accordance with contemporary community guidelines in human genomics data sharing [10] [11] [12].

FAIRification process:

Metadata, data, and supporting infrastructure are governed by the FAIR Data Principles (e.g., search engines). The requirements for discoverability and accessibility can be addressed at the metadata level. At the data level, tremendous effort is needed for interoperability and reuse. The diagram below demonstrates GO FAIR's FAIRification process, which emphasizes data but also incorporates metadata work:

Steps in the FAIRification process [13]:

1. Obtain access to non-FAIR data: acquire access to the data that will be FAIRified.
2. Analyze the information you've collected: Assess the information's content: What ideas are portrayed? What is the data's structure? How are the datasets linked to each other? Distinct data sets necessitate different analysis and evaluation approaches. If the dataset is stored in a relational database, the relational schema contains information on the dataset's structure, types (field names), cardinality, and so on.
3. Define the semantic model: create a semantic framework for the dataset that consistently, explicitly, and in a computer-actionable manner explains the meaning of objects and interactions in the dataset. A good semantic model should represent a shared understanding in a certain area and for a specific purpose. As a result, searching for existing models is a useful strategy. Several terms from ontologies and vocabularies are frequently used in semantic models.
4. Make data linkable: Using the semantic model, non-FAIR data may be turned into linkable data. This stage facilitates data integration with other data and systems by enabling interoperability and reuse. However, the user must determine whether or not this step is appropriate for the provided data. It's a wise idea for many sorts of data (e.g., structured data), but it might not be appropriate for others (e.g., the pixels or audio elements in images, audio data, and videos).

5. Assign licence: Although licence information is included in the metadata, it is a distinct step in the FAIRification process to emphasize its relevance. Even though the data is meant to be open access, the lack of clear licensing may prohibit others from exploiting it.
6. Define metadata for the dataset: Proper and rich metadata assist all elements of FAIR, as indicated by several of the FAIR principles.
7. Deploy FAIR data resource: deploy or publish the FAIRified data, together with the necessary information and a licence, so that browsers may index the metadata, and the data can be retrieved, even if credentials and authorization are required.

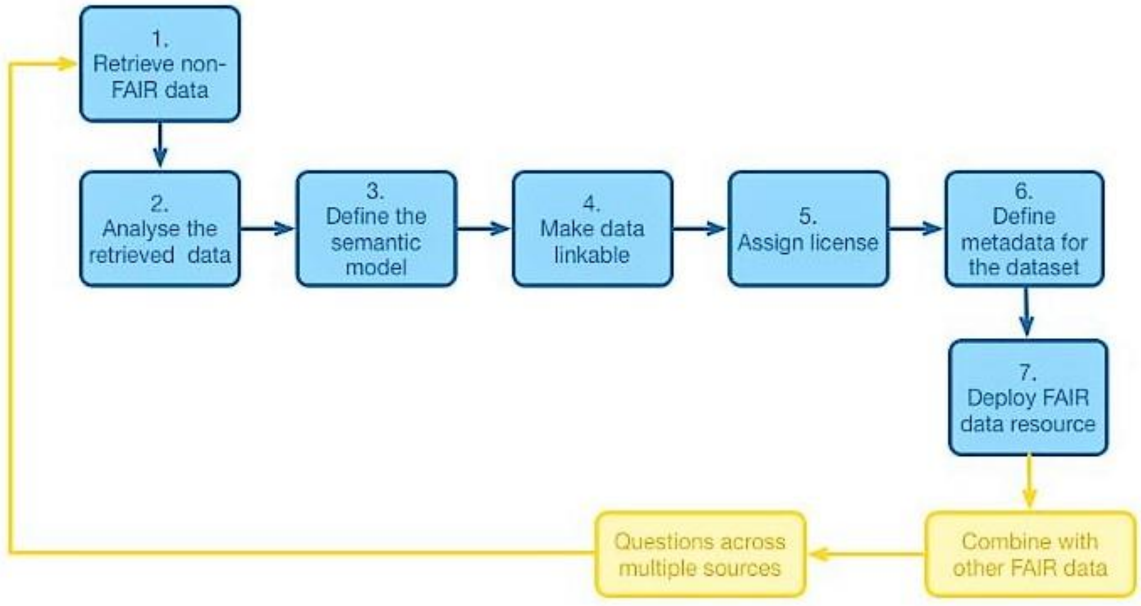

1. Retrieve non-FAIR data
2. Analyse the retrieved data
3. Define the semantic model
4. Make data linkable
5. Assign license
6. Define metadata for the dataset
7. Deploy FAIR data resource
Combine with other FAIR data
Questions across multiple sources

Fig.1. GO FAIR's FAIRification process [13]

An insight into Biomedical data:

Academic labs and research groups create massive volumes of biological multi-omics data. This data has a ton of potential for reuse, but it's dispersed among a number of different sources and isn't standardized. As a result, the data availability does not imply its utility, necessitating the development of a fast way to explore molecular data.

Data broadly fall into two categories:

1. Preclinical data: This type of data is generated/used by teams/organizations working on understanding a biological problem and driving milestones in the research and development process. Data sources can be cell-line (DepMap, LINCS, GEO), tissue or xenograft data. These teams are typically looking for hypothesis generation. For, e.g. Identifying a drug of interest and a patient segment who will benefit from the drug.
2. Clinical data: This type of data is typically generated/used by teams/organizations that are trying to understand the impact of a drug/treatment on a particular disease condition. These teams are typically approaching the end goal of finding validation datasets. For, e.g. Taking the drug to market

Data Sources

1. Public data
    1. Experimental data deposited by researchers - e.g. Gene Expression Omnibus
    2. Data generated/collated by a consortium - e.g.TCGA, DepMap, Human Cell Atlas
2. Proprietary data
    1. Experimental data generated/collated by for-profit/Not-for-profit groups - e.g. Genentech, FOUNDIN-PD.

Data Type

1. Using common multi-omics techniques - Transcriptomics, Metabolomics, Epigenomics, GWAS, Proteomics etc.
2. Using Phenotype studies (non-omics) - Flow cytometry, ELISA, Luminex etc

The biomedical data is not fair, and the FAIRification of this data for machine learning is a critical task.

Machine learning has the potential to unlock key values for organizations both from public and proprietary data.

Unfair data practices pose significant challenges to the research community.

The challenges arise because:

- Public databases are available but not usable.
- Proprietary data is often caught in team-level silos and hence has low interoperability and reusability.

Due to aforesaid reasons, a lot of time is consumed in cleaning the data and bringing it into a format which is suitable for the purpose. This reduces the efficiency of scientists before running an algorithm.

Polly can supplement those activities, saving ample time for the scientists by curating the data making it fair so that scientists can make the best use of their time.

POLLY:

Polly is a SaaS platform which provides access to ML-ready data. Elucidata's Polly is a cloud platform which gathers and stores public or premium biomedical data; the data is stored in curated form, which allows users to access and analyze highly curated, structured, harmonized data using either application (GUI) or coding (Polly Python). The data in Polly was initially stored in the form of data lakes which are now being converted to the highly curated OmixAtlases.

Being an AI-powered cloud-based platform, Polly enables customers to access curated, machine-learning-ready biomedical data from both public and private sources. ML algorithms are used to curate data from various repositories, databases, and proprietary datasets, ensuring that it is machine-actionable and analysis-ready.

By providing a library of robust, easy-to-customize bioinformatics workflows, Polly's cloud system facilitates seamless data analysis, visualization, and sharing. It establishes a one-of-a-kind, centralized environment that allows a varied group of biologists, bioinformaticians, and scientists to exchange and cooperate on workspaces, data, and insights.

Polly Supplements data acquisition, metadata curation data engineering, cloud infrastructure and analysis and reporting of the data.

Fair data is at the core of Polly:

Polly data lakes host multi-omics data such as transcriptomics (TCGA, GEO, Gtex), Proteomics (PRIDE), Metabolomics, Epigenomics and disease-associated data.

The IDEATE framework forms the backbone of the FAIRification process on Polly.

IDEATE Framework:

The IDEATE framework, detailed below, defines the key aspects of the data-driven problems that are solved through the platform, Polly.

Ingest

Polly receives data from regulated repositories, publications, and proprietary data via the Data Connectors. Data enrichment is done automatically with the ML-driven Curation Infrastructure. The Analysis Pipeline instigates end-to-end data pre-processing.

Discover

The degree of data curation allows for both point-and-click filtering and complex code-based searching over the complete OmixAtlases data collection. Polly Libraries enables you to run complicated queries over various datasets, samples, and characteristics. The curation process streamlines data search by providing comprehensive and consistent metadata annotations with scientific context and excellent accuracy.

Enrich & Analyze

Polly's strong computational infrastructure can analyze ML-ready data by launching apps and running Notebooks. Polly Libraries provide a lot of freedom when it comes to comprehending data using code, allowing for holistic data analysis. Scientists can access and analyze their data from the platform through their own computational infrastructure.

Communicate

Data analysis is facilitated by built-in applications and configurable visualization dashboards. Reports created for the data, code, analysis, and outcomes with a single click that can be shared with the team. Workspaces on Polly allow customers to securely organize and manage the data.

Technology for data ingestion, pre-processing and curation

1. Data Engineering and Harmonization

   Every dataset on Polly undergoes two key steps that make it machine-readable and ML-ready

   - *Data engineering involves consistent data ingestion and storage to provide data formats that allow easy access to actionable data.*
   - Metadata Harmonization, which follows standardized ontologies for annotation at the dataset and sample level to increase the findability of data for reuse.

2. Polly Connectors
   - Before Polly data resides in the organization, the depositors seem fit for their own use; reuse and consistency of metadata is rarely the goal.
   - On Polly, data curation is the goal which is performed by harmonizing multiple file formats into a consistent organization on a single data infrastructure.

3. PollyBERT
   - A lot of metadata is semistructured or follows the schema defined by each depositor, which results in incomplete or inconsistent metadata. Omics data are unstructured and hence not in analysis-ready format (.csv, gct, Excel, H5ad format).
   - Through the curation pipeline, data is harmonized with the metadata using ontologies and saved the data in accessible formats either as gct files which support a lot of omics and non-omics data or as h5ad files that support more large and complex data like single-cell RNAseq to save the data along with metadata in the same file.

4. Curation:
   - Process of transforming the data into a consistent and machine-readable first and making it accessible on a platform guided by the FAIR principle. The process of curation involves both people and advanced ML.
   - The public and proprietary data is streamlined into one consistent schema; the data sets on Polly are stored in GCT format, which allows for storing sample and molecular level metadata in a single file. Single-cell data are stored in H5ad format. Ontology-based mapping of text related to disease, cell type, cell line, tissue, and drug allows data to be curated at scale.

Curation at Scale:

Polly currently hosts more than seven lacs datasets which consist of public, premium and proprietary datasets. More than 20 data types are available on Polly from 32 data sources. Datatypes available on Polly are namely Single-cell, metabolomics, proteomics, drug screening, Gene effect and mutation etc. (Fig.2.).

Four million samples from data sources are hosted on Polly, which are curated before they are made available on Polly. The data comprises more than 300 organisms, and out of it, 80% belongs to humans.

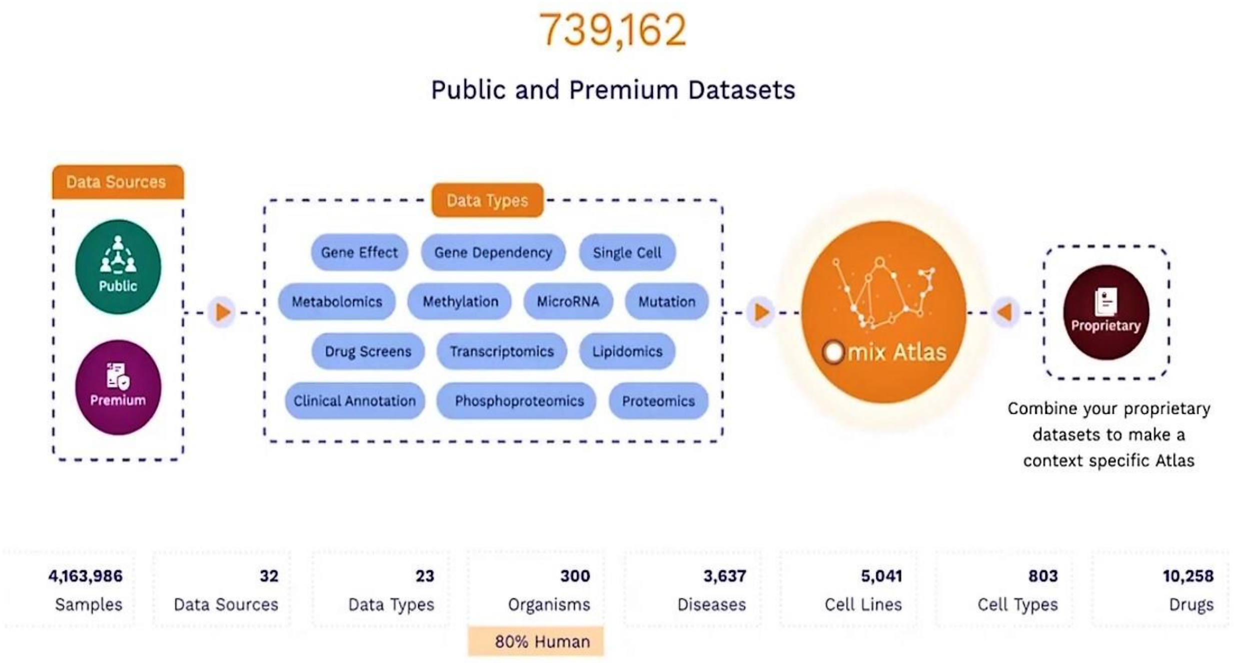


Fig.2. Data types on Polly

Metadata annotation:

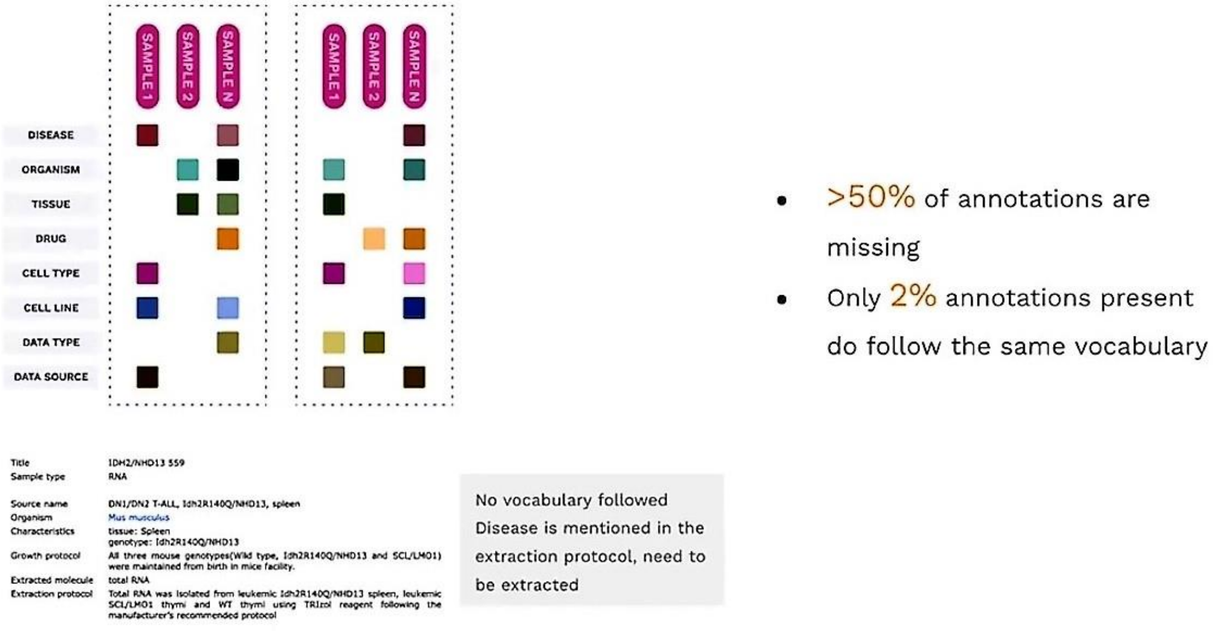
Metadata annotation is an indispensable part of data curation to bring in the quality of the work process.

In simpler language, it can be defined as annotating different kinds of metadata in different data sets using the same molecular identifiers and tagging each sample with similar information.

Metadata at source

The metadata in the source is in poor shape. More than 50 per cent of the annotations are missing, and only 2% of annotations present follow the same vocabulary.

The figure (Fig.3.1) exhibits the annotation discrepancy as the disease name is not explicitly mentioned. The disease name is T-ALL (T- cell acute Lymphoblastic Leukemia) and is expressed in a critical manner under the source name. As the name of the disease is not mentioned, it becomes almost impossible for scientists to reuse the data despite it being available and accessible in public databases. Until we have a very close look at each and every piece of information available with the data set identifying the name of the disease is Sisyphean.



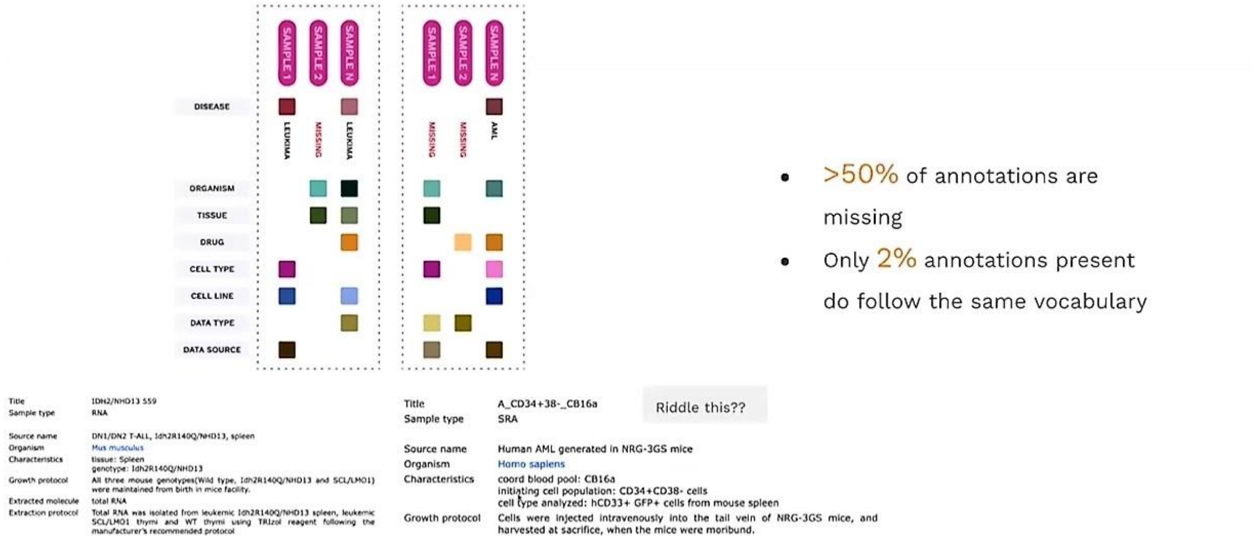
- >50% of annotations are missing
- Only 2% annotations present do follow the same vocabulary

Title	IDH2/NHD13 559
Sample type	RNA
Source name	DN1/DN2 T-ALL, Idh2R140Q/NHD13, spleen
Organism	Mus musculus
Characteristics	tissue: Spleen genotype: Idh2R140Q/NHD13
Growth protocol	All three mouse genotypes(Wild type, Idh2R140Q/NHD13 and SCL/LMO1) were maintained from birth in mice facility.
Extracted molecule	total RNA
Extraction protocol	Total RNA was isolated from leukemic Idh2R140Q/NHD13 spleen, leukemic SCL/LMO1 thymi and WT thymi using TRIzol reagent following the manufacturer's recommended protocol

No vocabulary followed
Disease is mentioned in the extraction protocol, need to be extracted

Fig.3.1. Metadata at source

Similarly, in the other figure (fig.3.2), it is difficult to identify the disease name Acute myeloid leukemia as it is present under the source name, making it almost impossible to recognize the disease until one glances closer at the source name.



- >50% of annotations are missing
- Only 2% annotations present do follow the same vocabulary

Title	IDH2/NHD13 559
Sample type	RNA
Source name	DN1/DN2 T-ALL, Idh2R140Q/NHD13, spleen
Organism	Mus musculus
Characteristics	tissue: Spleen genotype: Idh2R140Q/NHD13
Growth protocol	All three mouse genotypes(Wild type, Idh2R140Q/NHD13 and SCL/LMO1) were maintained from birth in mice facility.
Extracted molecule	total RNA
Extraction protocol	Total RNA was isolated from leukemic Idh2R140Q/NHD13 spleen, leukemic SCL/LMO1 thymi and WT thymi using TRIzol reagent following the manufacturer's recommended protocol

Title	A_CD34+38-_CB16a
Sample type	SRA
Source name	Human AML generated in NRG-3GS mice
Organism	Homo sapiens
Characteristics	coord blood pool: CB16a initiating cell population: CD34+CD38- cells cell type analyzed: hCD33+ GFP+ cells from mouse spleen
Growth protocol	Cells were injected intravenously into the tail vein of NRG-3GS mice, and harvested at sacrifice, when the mice were moribund.

Riddle this??

Fig.3.2. Metadata at sources

At Polly, the scale of curation is huge, with approximately 4.1 million curated samples and 1.2 million manually curated labels present on Polly today. Also, it is a norm to add five fields every quarter to Polly.

Fig.5. denotes the difference between annotations between the sources of data and Polly. It is evident from the figure that Polly has <1% of missing annotations, and annotations follow the same vocabulary/ ontologies throughout the platform, whereas the non-curated data at data sources for, e.g. The Human Protein Atlas, TCGA, GEO has only 2% annotations following the same vocabulary and >50% annotations are missing.

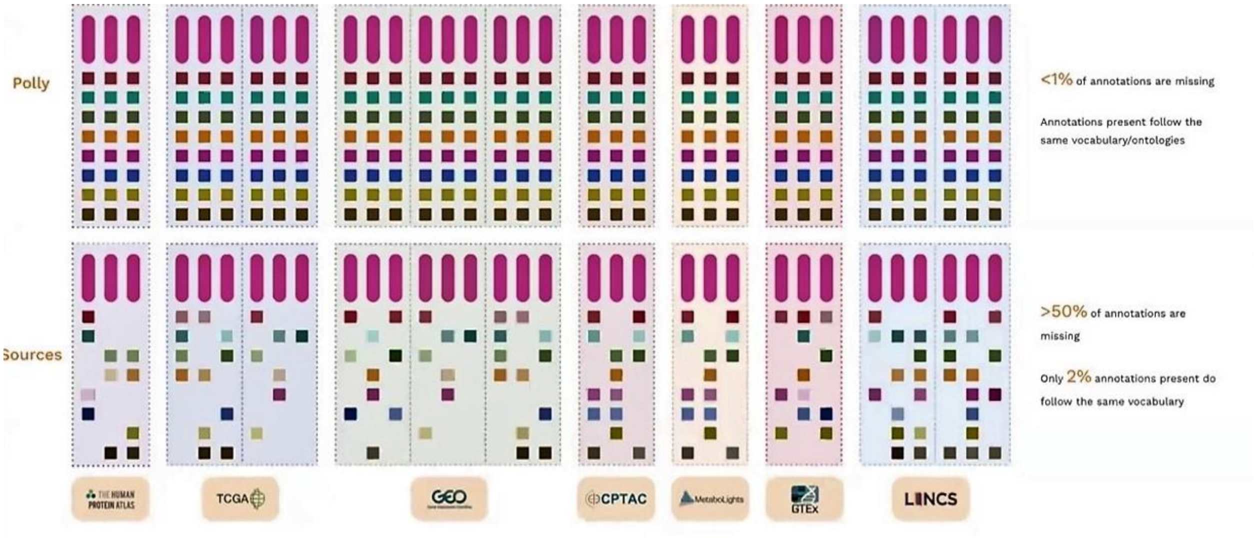
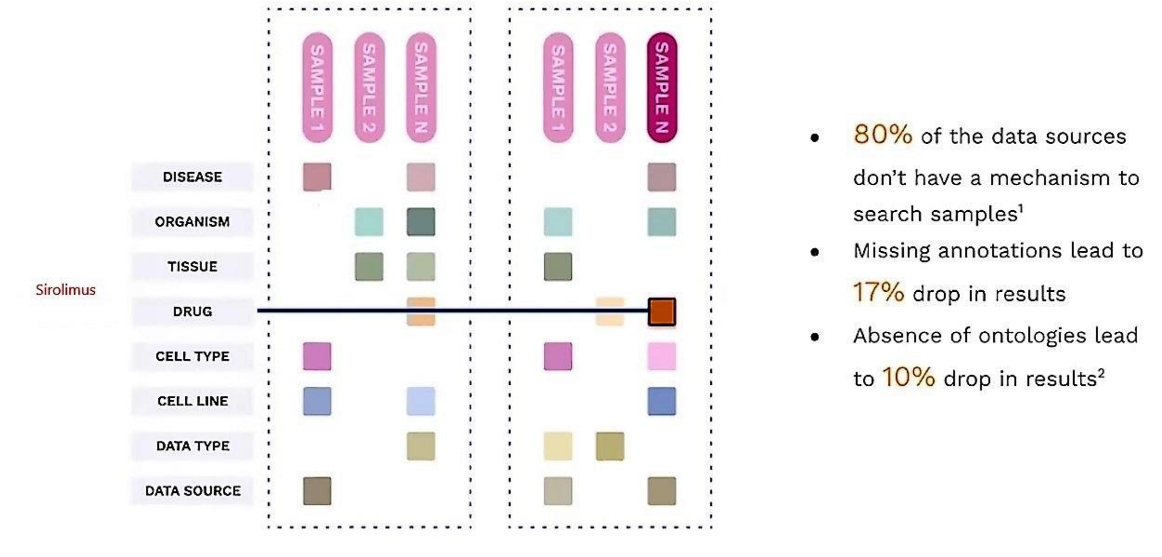

Polly
Sources
<1% of annotations are missing
Annotations present follow the same vocabulary/ontologies
>50% of annotations are missing
Only 2% annotations present do follow the same vocabulary
THE HUMAN PROTEIN ATLAS
TCGA
GEO
CPTAC
MetaboLights
GTEx
LINCS

Fig.4. Polly Vs Sources

Need for curation at scale:

It can be stipulated from the figure that 80% of the data sources don't have a mechanism to search samples. Due to missing annotations, there is again a drop of 17% in results for samples, and further lack of standard ontologies reduces 10% reduction in sample search. Ultimately the quantity of data retrieved is minuscule, and the quality of data is compromised. This reduces the scope of FAIR data.

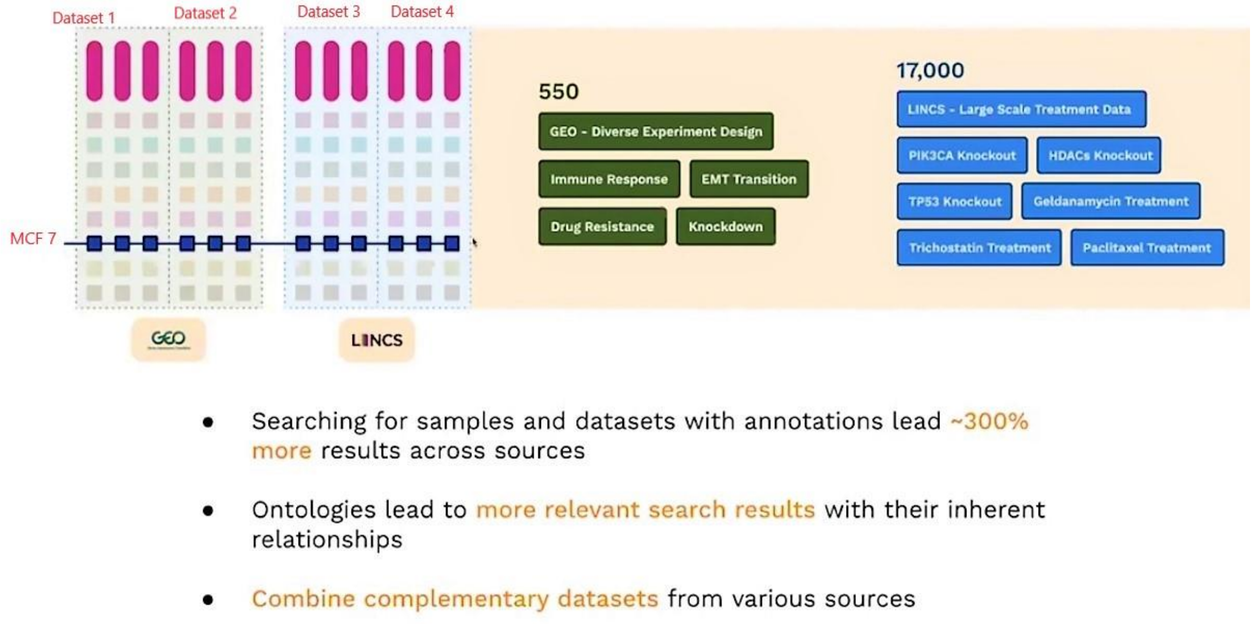
For instance, in the figure (Fig.5), there might be a lot of data available for the Drug Rapamycin, also known as Sirolimus; however, due to a lack of standard ontology, the drug samples might be present with different names making it difficult to find all the samples for the drug.



- 80% of the data sources don't have a mechanism to search samples[1]
- Missing annotations lead to 17% drop in results
- Absence of ontologies lead to 10% drop in results[2]

Fig.5. Missing annotations

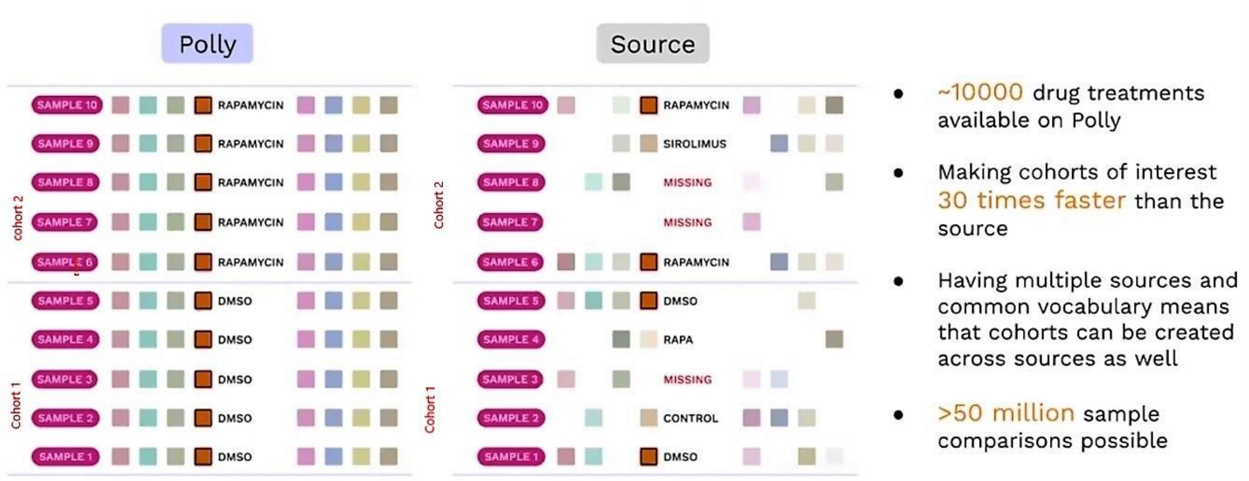
At Polly, when one searches for a dataset, the results exhibit data not only from one source but also from other relevant sources, increasing the number of datasets with annotation by 300% more which in many cases may be complementary that can help the researcher to gain a deeper dive into the data. For instance, GEO contains data relevant to the immune response in diverse design experiments, and LINCS have data relevant to treatment. The information can hence be combined to answer a lot of biological questions (Fig.6.)



- Searching for samples and datasets with annotations lead ~300% more results across sources
- Ontologies lead to more relevant search results with their inherent relationships
- Combine complementary datasets from various sources

Fig.6.Searching for Datasets at Polly

Another example is if a researcher wants to analyze Rapamycin treatment across various cell lines or tissues. The researcher might have to learn different sources and standardize the ontologies themselves to create the cohorts for different samples, but at Polly, one can find the cohorts of Rapamycin DMSO and controls in a prefabricated from across various sources on Polly (Fig.7.). This increases the efficiency of the researchers and makes cohorts of interest 30 times faster than the source.



- ~10000 drug treatments available on Polly
- Making cohorts of interest 30 times faster than the source
- Having multiple sources and common vocabulary means that cohorts can be created across sources as well
- >50 million sample comparisons possible

Fig.7.Analysis on Polly Vs Sources: Rapamycin treatment

Steps in Curation:

Data is collected from various sources (private/ proprietary) and then annotated weakly by ML models, after which expert annotation is carried out by bioinformaticians and scientists. This data is then fed to the ML model, which learns and curates better in the next annotation. As the number of iterations increases, the ML model efficiency also increases. The steps involved are:

1. Data download:

Bio-medical data from a multitude of sources such as GEO, GDC, TCGA, LINCS, *ImmPort*, and publications. Also, the metadata associated with these data may be available along with the data or separately. These will be downloaded and stored.

2. Data audit and ingestion:

Data auditing is carried out to assess how a dataset is fit for a given purpose. Ingestion gathers data and brings it into a data processing system where it can be stored, analyzed, and accessed. When the data source is a public repository, the data type and the file formats which will be present in the repository are known. In this case, a data audit is carried out first and then the ingestion process can take place. On the other hand, if the data to be processed is proprietary data, it could be ingested first to organize it before it can get into an auditable form.

3. Curation:

The curation process can be manual, automatic, or a combination.

Manual curation methodology:

1. Double-blinded curation - In this approach, two different curators are assigned with the same data and the same guidelines to curate independently.
2. Consensus - After independent curation, the curated data will be compared. If the data matches 100%, it will be sent for expert review. If the matching score was less than 100%, the data would be sent back to curators for post-consensus discussion.
3. Post Consensus Discussion - For each data, the curators will discuss the resources and the values. After mutual agreement, the value for each field will be finalized.
4. Expert Review - The values in each field for both standard information and experimental information will be reviewed by the expert team

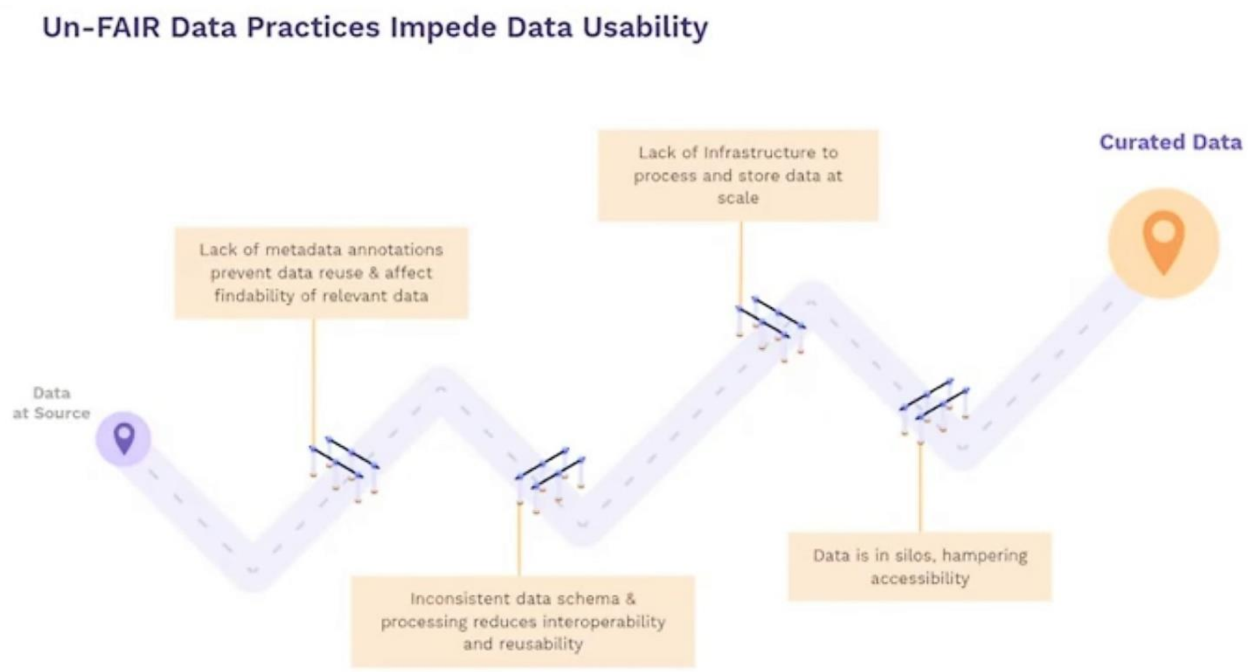
ML-based curation methodology:

To support and accelerate the manual curation process, Polly has Polly Bert, which is one of the most optimized models in the biomedical industry, trained on 17 billion words and 660 million parameters. This model helps in weakly annotating the datasets according to a given set of rules and also actively learns from each processed dataset.

Breaking barriers to data usability (Fig.9.)

Table 1. Pain Points in biomedical research and their solutions

Pain Points	Solution
Lack of metadata annotations prevents data reuse and affect the findability of relevant data	Metadata harmonization by mapping metadata to known biomedical ontologies
Inconsistent data schema and processing reduce interoperability and reusability	Data streamlined in one consistent schema and processed with state of art pipelines
Lack of infrastructure to process raw data at scale	Scalable cloud platform to store and process data at scale
Data is in silos, hampering accessibility	Access to centralized atlas of curated data on Polly


Un-FAIR Data Practices Impede Data Usability

Lack of metadata annotations prevent data reuse & affect findability of relevant data

Inconsistent data schema & processing reduces interoperability and reusability

Lack of Infrastructure to process and store data at scale

Data is in silos, hampering accessibility

Fig.8. Barriers in Data usability

Case Study to highlight the importance of the Fairification process- Roche [14]

Roche's main goal was to use FAIR and shareable data to accentuate the development of new pharmaceutical therapies.

The objective was to not only FAIRify legacy datasets in various Therapeutic Areas (TA) but also to build mechanisms for future research's potential FAIRification.

The following were part of the first strategic plan:

1. Identify and develop use cases, as well as research studies that are relevant. Age and applicability of a dataset, and the significance of a study for fostering translational research, all are factors to explore.

   Find the datasets, relevant documents, and research groups.

1. Restriction of access based on Study Informed Consent Forms and study status ensures data protection.
2. Determine which sections of the data require FAIRification. Or only the data that'll be useful in drug discovery and development?
3. Design the FAIRification process and resources. Using the Clinical Data Interchange Standards Consortium and sponsor extensions, use reference data to harmonize and standardize the data set. The FAIR principles are completely implemented in Roche's reference data, including the use of Unified Resource Identifiers (URIs) for data and metadata.
4. To generate a FAIR version, execute mapping and transformation processes on the research.
5. Perform thorough quality assurance and evaluate FAIRness in line with curation norms.
6. Value is determined through close collaboration with data scientists.

Outcomes:

Outcomes from Roche's data FAIRification process

Roche effectively used FAIRification in four of their therapeutic fields, including ophthalmology, autism, asthma, and COPD. FAIRification strengthened its procedures to better curation and management of data and metadata as a consequence of the project.

The three important lessons that Roche noticed were -:

- Because of the unequal adoption of data management standards, Roche noticed that data comprised of different models over time.
- Due to the company's vast growth, system evolution, and personnel leaving or moving on to other areas, Roche noticed various challenges linked to data Findability and Accessibility.
- The third most important obstacle was a cultural one: an unawareness within individuals of the merits of FAIR data.

FAIRification culture at Elucidata:

Elucidata understands that FAIR Implementation for clinical data necessitates a cultural shift to a data-centric mindset that is followed by everyone as best practice. Polly is designed to combine end-to-end curation workflows to allow data FAIRification at scale, hence improving the FAIRness of healthcare data and metadata.

Challenges in Implementation of Fair data [15]:

- The cost element involves the initial cost of modifying current data to meet data standards, as well as the historical expenditure in old systems and the cost of data loss during the transition. While there'd be some upfront expenditures associated with becoming FAIR compliant, a study conducted by the European Commission assessed the minimum yearly cost of scientific fields without FAIR data to be €10.2 billion throughout the EU. FAIR is not only cost-effective to adopt, but it also provides long-term financial rewards.

- Culture shock: According to research, the second most difficult obstacle for life science organizations seeking to establish data-driven and digital competencies is culture. Cultural barriers might include a lack of understanding of data principles, misconceptions about the difficulty of applying FAIR, and academics' attachment to old systems. These may be readily overcome by further education about the need for data standards, both within the industry and at the university level.

- The compliance barrier: Changing procedures can be difficult due to the extensive regulation that exists throughout scientific sectors, ranging from health and safety rules in chemicals to approval cycles in life sciences. FAIR data standards, on the other hand, are advantageous from a regulatory standpoint, guaranteeing that. While some people believe that applying FAIR requires all data to be available, this is not the case — especially when it comes to patient confidentiality, legal sensitivity, or economic interests. 'Closed' data, on the other hand, should not be excused from being FAIR just because it isn't publically accessible.

Way forward with FAIR data:

- Enhances the rate of discovery

When datasets are openly accessible, they may be easily accessed and utilized to build a more comprehensive view of a particular topic, or they can be studied by data mining tools to find connections that the creators of the original data were oblivious of.

- Ensuring that no major breakthroughs are missed.

Any given dataset may be used or analyzed in a variety of different ways. To someone with a different viewpoint or analytical method, what seems to be noise to one individual might be a significant finding to someone else.

- Enhances the scientific and academic record's integrity

Researchers can double-check one other's research and confirm that conclusions are supported by scientific evidence when the data that underpins findings is accessible.

- Many in the research community are beginning to see open and FAIR data as a crucial aspect of today's research business.

Institutions participating in the research process, from research funders to publishers, are beginning to demand that, at the absolute least, the data that underpins publications be made readily accessible.

A FAIR future for all?

While it is not required for organizations or individuals to use data management principles, it is strongly encouraged in today's rapidly changing scientific field. We're at the beginning of AI-driven science, which has the potential to address some of the world's most pressing scientific issues, and data will be at the focus of it all.

We have the opportunity now to work together and develop guidelines for our data to advance research. The alternative is to wait for the stage when technology compels organizations to format and standardize all their data. Then, the companies without FAIR standards will be one step behind, racing to meet the basic requirements needed to derive value from their data.

We now have the possibility to collaborate and set criteria for our data in order to further study. Another option is to wait until technology forces organizations to format and standardize all of their data. Companies that do not adhere to FAIR principles will fall behind, racing to achieve the minimum criteria for extracting value from their data.

References

1. van Vlijmen H, Mons A, Waalkens A, Franke W, Baak A, Ruiter G, Kirkpatrick C, da Silva Santos LO, Meerman B, Jellema R, Arts D. The need of industry to go FAIR. Data Intelligence. 2020 Jan 1;2(1-2):276-84.
2. Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, Mellino G, Harrow I, Smith I, Taubert J, van Bochove K. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug discovery today. 2019 Apr 1;24(4):933-8.
3. Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LO, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use. 2017 Jan 1;37(1):49-56.
4. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016 Mar 15;3(1):1-9.
5. Fair cookbook [Internet]. FAIR Cookbook. The FAIR Cookbook; 2020 [cited 2022Jun14]. Available from: https://faircookbook.elixir-europe.org/content/home.html
6. Fair cookbook [Internet]. FAIR Cookbook. The FAIR Cookbook; 2020 [cited 2022Jun14]. Available from: https://faircookbook.elixir-europe.org/content/recipes/introduction/metadata-fair.html
7. Munafò, M. R., Nosek, B. A., Bishop, D., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. (2017). A manifesto for reproducible science. Nature human behaviour, 1, 0021. https://doi.org/10.1038/s41562-016-0021
8. Fecher B, Friesike S, Hebing M. What drives academic data sharing?. PloS one. 2015 Feb 25;10(2):e0118053.Boeckhout, M., Zielhuis, G. A., & Bredenoord, A. L. (2018). The FAIR guiding principles for data stewardship: fair enough?. European journal of human genetics, 26(7), 931-936.
9. European Commission, Directorate-General for Research and Innovation, *Open innovation, open science, open to the world : a vision for Europe*. Publications Office; 2016. Available from: doi/10.2777/061652

10. Boeckhout M, Zielhuis GA, Bredenoord AL. The FAIR guiding principles for data stewardship: fair enough?. European journal of human genetics. 2018 Jul;26(7):931-6.
11. Lezaun J, Montgomery CM. The pharmaceutical commons: Sharing and exclusion in global health drug development. Science, Technology, & Human Values. 2015 Jan;40(1):3-29.
12. Knoppers BM. Framework for responsible sharing of genomic and health-related data. The HUGO journal. 2014 Dec;8(1):1-6.
13. GO fair Initiative [Internet]. Go FAIR. 2018 [cited 2022Jun14]. Available from: https://www.go-fair.org/fair-principles/
14. FAIRification of clinical trial data – Roche [Internet]. 2017 [cited 2022Jun14]. Available from https://fairtoolkit.pistoiaalliance.org/use-cases/fairification-of-clinical-trial-data-roche/
15. Alharbi E, Gadiya Y, Henderson D, Zaliani A, Delfin-Rossaro A, Cambon-Thomsen A, Kohler M, Witt G, Welter D, Juty N, Jay C. Selection of data sets for FAIRification in drug discovery and development: Which, why, and how?. Drug discovery today. 2022 May 17.
16. Chen B, Butte A. Leveraging big data to transform target selection and drug discovery. Clinical Pharmacology & Therapeutics. 2016 Mar;99(3):285-97.
17. Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug discovery today. 2018 Aug 1;23(8):1538-46.
18. Brown N, Cambruzzi J, Cox PJ, Davies M, Dunbar J, Plumbley D, Sellwood MA, Sim A, Williams-Jones BI, Zwierzyna M, Sheppard DW. Big data in drug discovery. Progress in medicinal chemistry. 2018 Jan 1;57:277-356.
19. Wilkinson MD, Sansone SA, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. A design framework and exemplar metrics for FAIRness. Scientific data. 2018 Jun 26;5(1):1-4.
20. Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, Mellino G, Harrow I, Smith I, Taubert J, van Bochove K. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug discovery today. 2019 Apr 1;24(4):933-8.
21. European Commission, Directorate-General for Research and Innovation, *Cost-benefit analysis for FAIR research data : cost of not having FAIR research data*. Publications Office; 2019. Available from: doi/10.2777/02999