Agency for Science, Technology and Research
SINGAPORE

Biomedical Data Hub

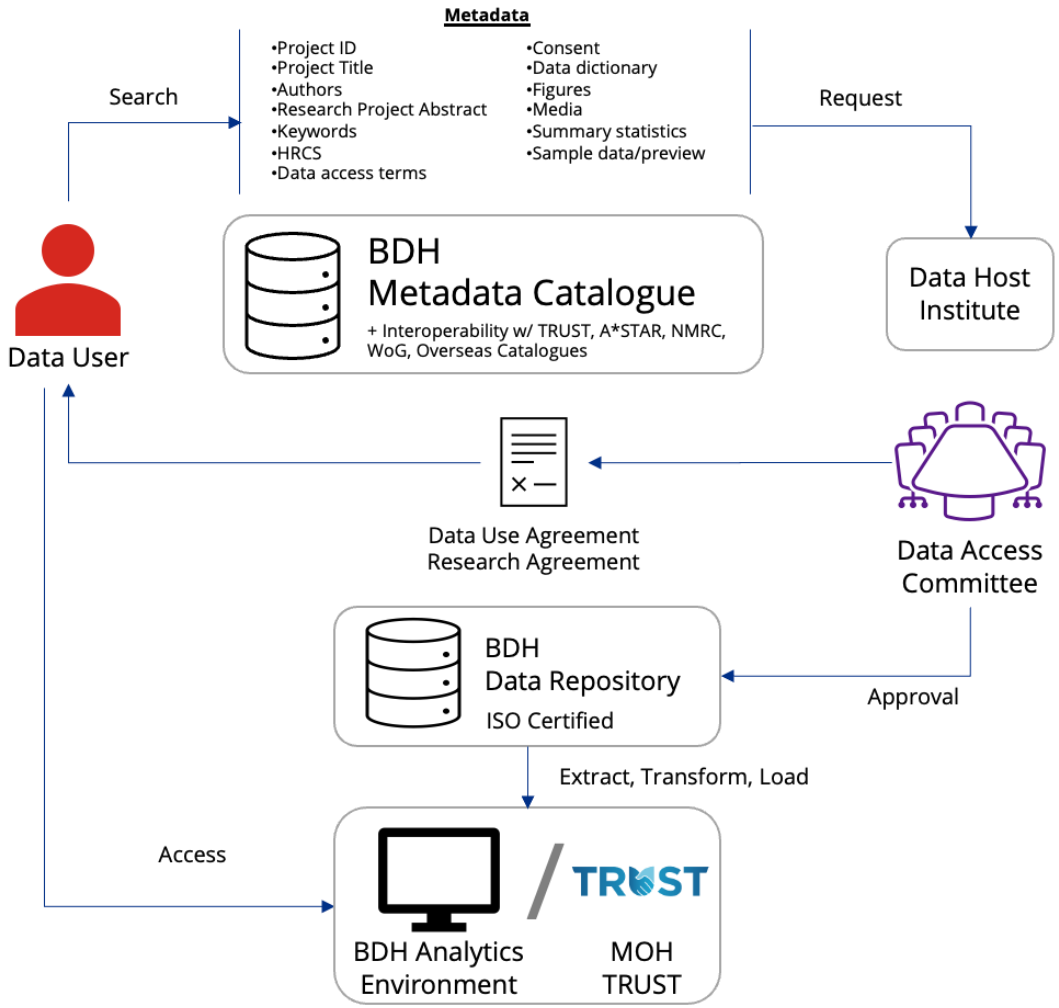
BDH Satellite Roadshow Hosted by NCO

David Vu
Programme Director, BDH

9 February 2026

BDH in Context of the Human Health and Potential Data Ecosystem in Singapore

Biomedical Data Hub helps makes Singapore research data Findable, Accessible, Interoperable and Reusable (F.A.I.R.)



Use Cases

Discovery

Search for relevant data and quickly assess its utility

Governance

BDH tech infra ensures data owners retain control over access and use

Analysis

Data is accessed and analysed in a separate and secure environment

Capabilities

Catalogue

User searchable metadata catalogue *(Coming Soon)*

Repository

ISO certified secure multi-modal data repository

Analytical Node

Trusted research environment (TRE) matched to compute and linkable to clinical data[1]

National Data Curation Team

Coordinates data standards working groups and a federated resource pool of data curators

Project Management

Experienced bioinformatians to help facilitate access and use of data for research and innovation

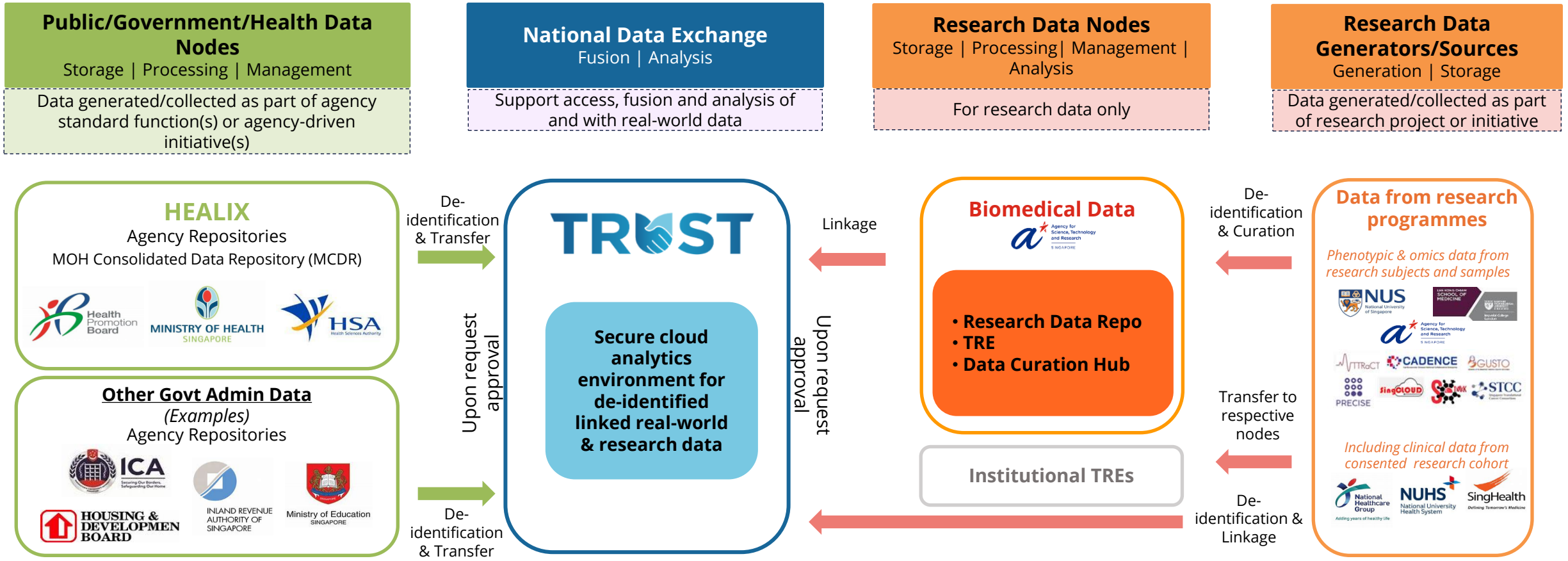
Data Types and Data Sets

- Deidentified research data
- Multiple clinical domains (Cardiology, oncology initially)
- Multi-modal data: Tabular, free-text, imaging, omics, ECG

(1) Clinical data linkage via MOH TRUST

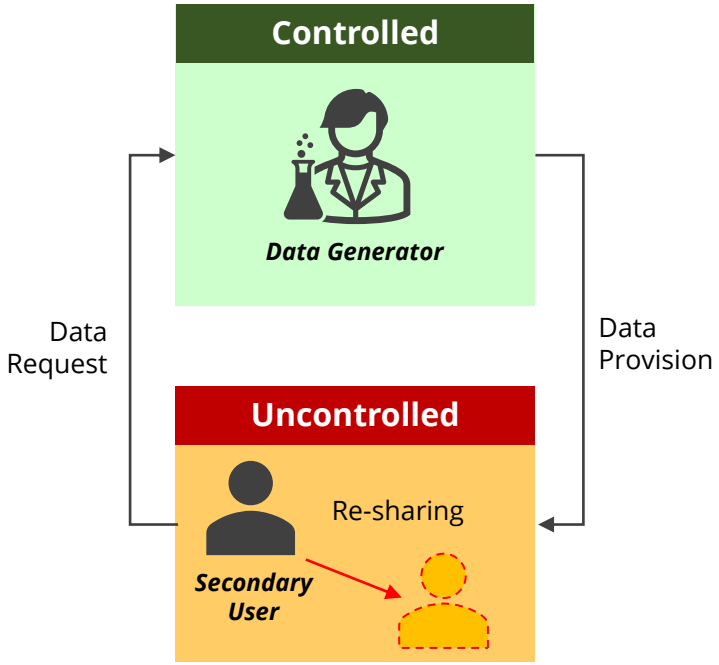
BDH and TRUST Act Together as Trusted Research Environments for Real World and Research Data

BDH supports TRUST as a secure processing node, analytics environment and repository for de-identified research data



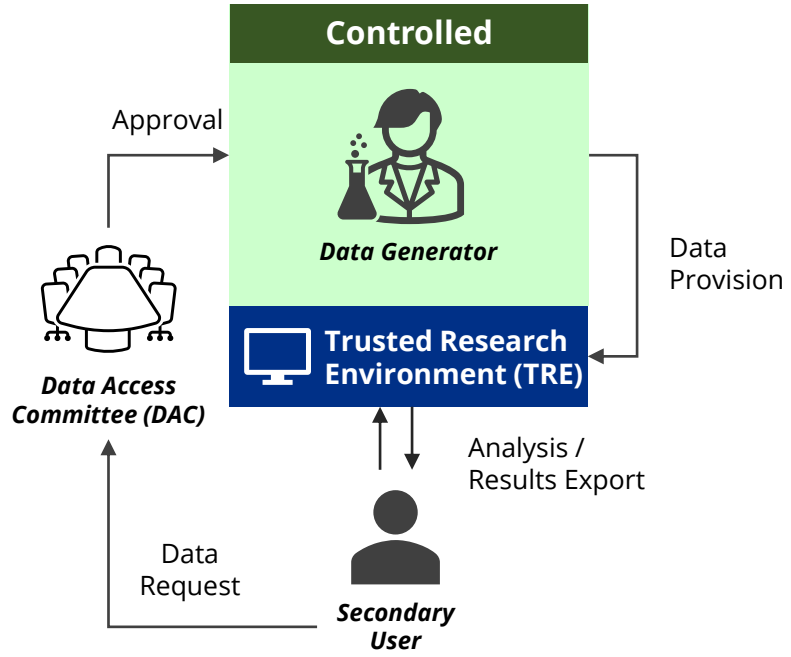
TREs Address Issues with Research Data Sharing… BDH Gives Data Owners a Central Option to Avoid Over Proliferation

The Good 'ol Days



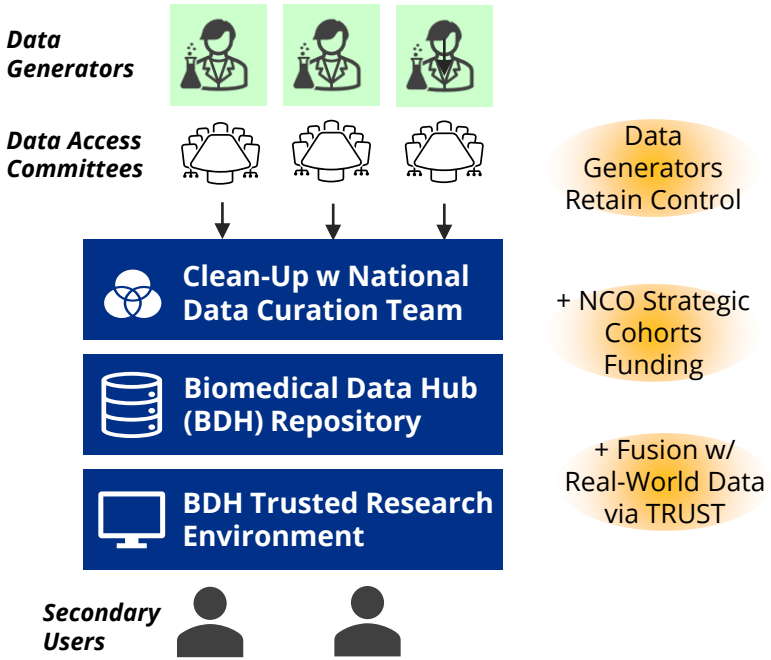
- ✓ *Laissez-faire* approach to data sharing
- 👎 Lack of formal governance
- 👎 Loss of control over data
- 👎 Data not discoverable or standardized

Trusted Research Environments



- ✓ Data stays in secure TRE
- ✓ Governance ensures public trust
- 👎 Proliferation of TREs = Data silos
- 👎 Expensive to implement and operate
- 👎 Variable service levels

SG Biomedical Data Hub



- ✓ ISO certified security and governance
- ✓ Data discoverability via catalogs
- ✓ Able to fuse research datasets
- ✓ Direct linkage with TRUST
- ✓ Economies of scale

Challenges We've Heard from Data Generators and Users

Data Generators

- Increasing perils associated with data retention: **security**
- Uncertain **governance** and data ownership
- **Data retention obligations** extend beyond grant end dates
- Time and expertise required to stay abreast of **evolving data standards**

Data Users

- Difficult to **discover** available cohorts and datasets
- Lack of visibility to **patient consent** and other metadata
- Complex **governance** processes to approve and contract for access
- Access to tools to **analyze** data in trusted research environments

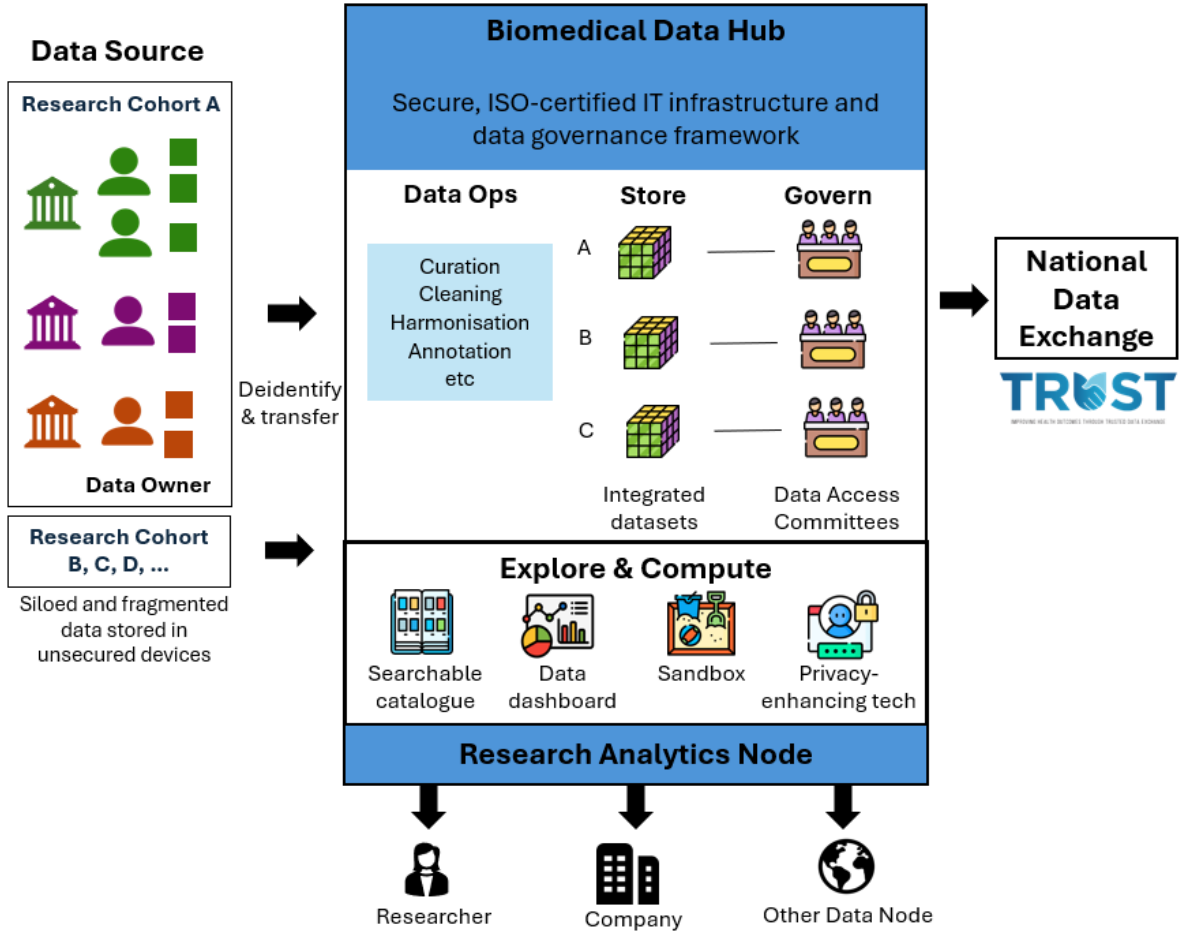
Biomedical Data Hub as a National Platform for Strategic Research Data

DIY Data Management Becoming Increasingly Challenging For Individual Researchers

Capabilities Required to Build a Data Repository on Analytics Environment

Governance	IT Infra Structure	Data Management	Business Management
Quality Management	Capacity Management (Storage)	Data Lifecycle Management	Project and Program Management
Risk Management	Compute infrastructure	Identity and Access Management	Product Management
Data Governance (Data Access Committee)	Business Continuity (Failover)	Output Management	Financial Management / Grant Management
User Accreditation	Software Development	Information Search and Discovery	Procurement
Training Delivery and Management	Solutions Architecture	Data Classification	Stakeholder Management
Data Privacy and Compliance (ISO Certification)	Systems Administration	Metadata Management	Public engagement
	End User Computing	Metadata Search and Discovery Application	Compliance and Regulatory Affairs
	Cybersecurity	Data Archiving	Legal

Biomedical Data Hub Serves as a Service Enhanced Repository and Analytics Node for Singapore's Strategic Research Data Assets



Vision:

Two Biomedical Data Repositories…

Bioinformatics Institute, A*STAR BII — *BII: Biomed DAR*

Genome Institute of Singapore, GIS — *GIS: RAPTOR*

…One governance and services layer

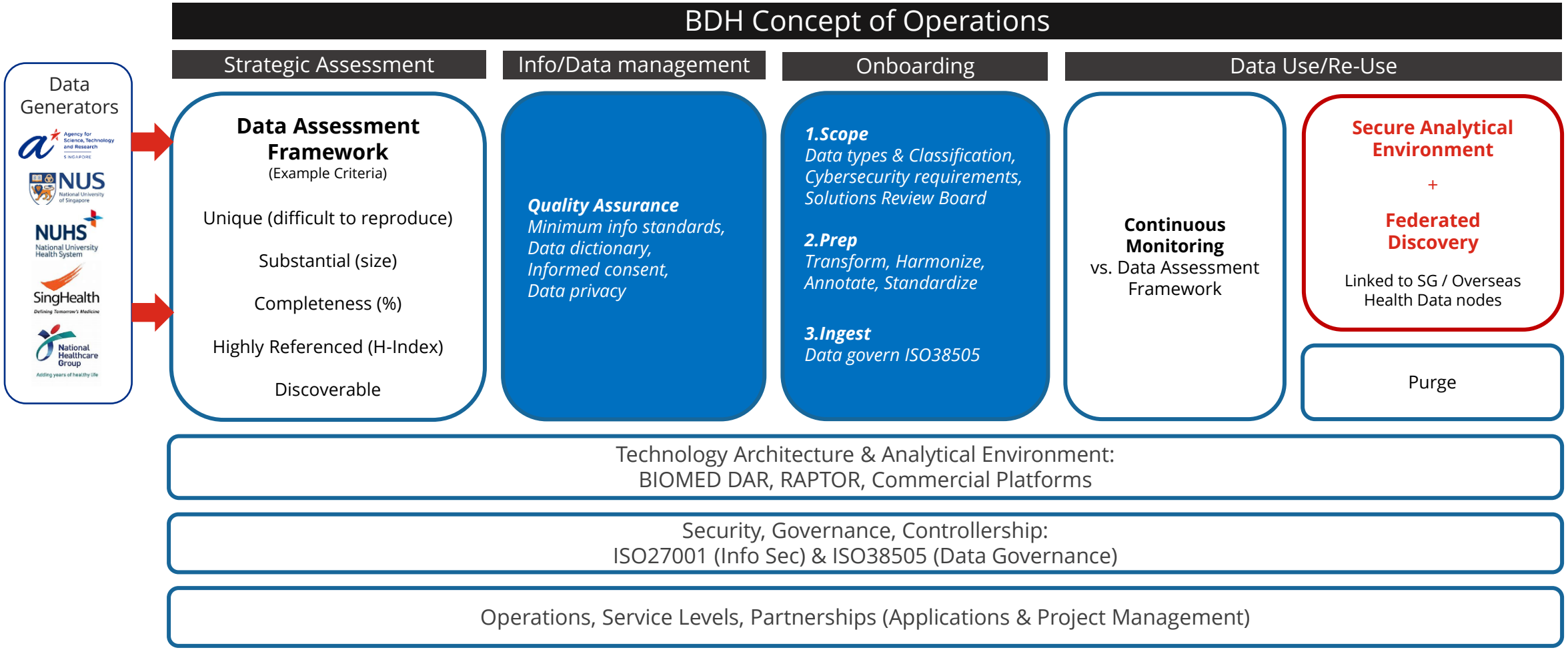
- ISO27001 (Info Sec), ISO38505 (Data Gov)
- Common SOPs and Governance

BDH Services

Comprehensive data management solution:

- Secure repository for data discovery or retention
- Data lifecycle management
- Data governance
- Data operations (e.g. curation and standardisation)
- Secure collaboration (e.g. privacy-enhancing technologies)

Biomedical Data Hub Concept of Operations Focuses on Data Governance, and Secondary Use



BDH will Facilitate Discoverability, Governance and Analysis, Encouraging Secondary Use of Research Data

1 Discovery

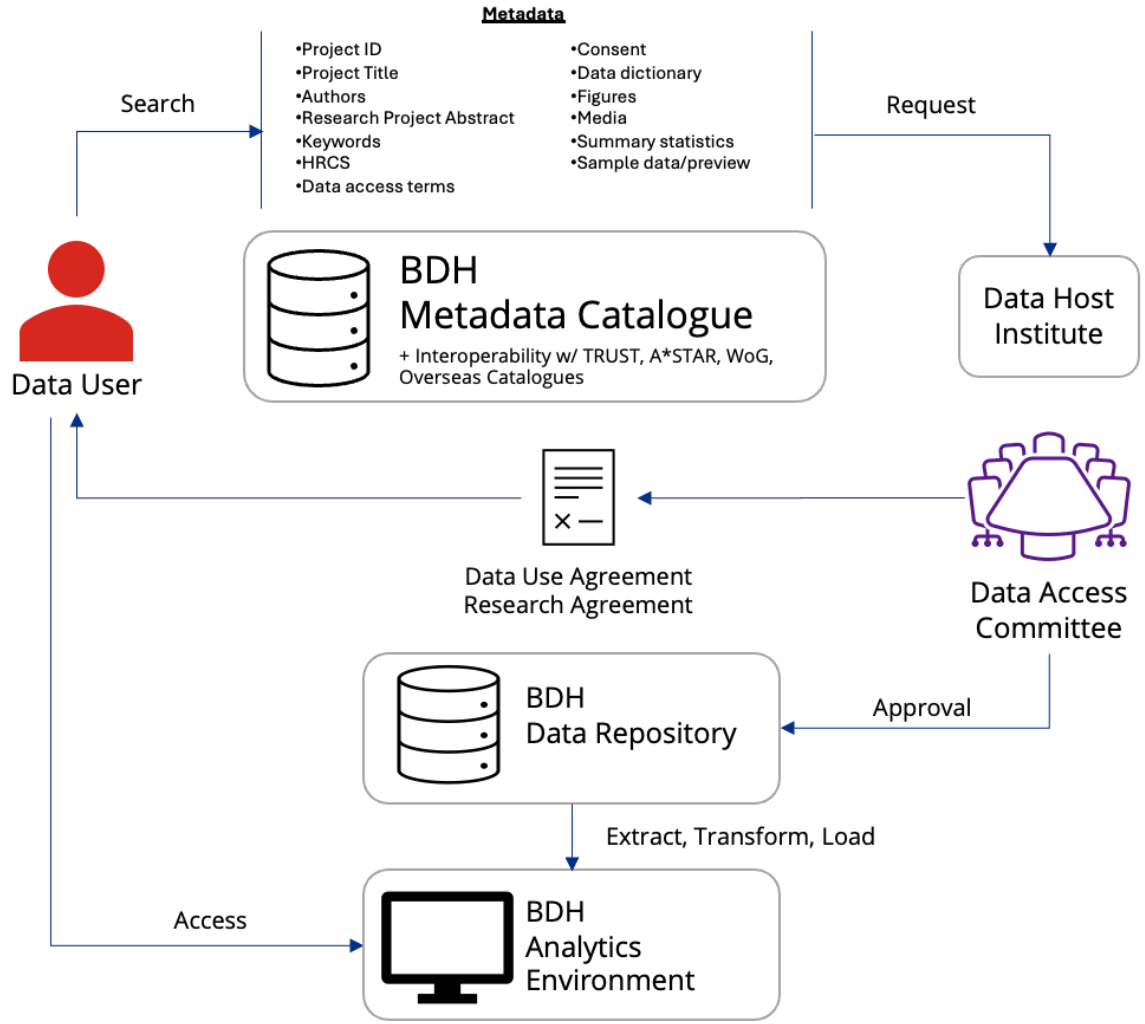
User finds dataset in data catalog and initiates request or is immediately granted access to public domain data

2 Governance

DAC at data host institute reviews data access requests and initiates governance / contracting processes

3 Access

Upon Approval BDH grants access in a separate and secure environment under the terms of the data use agreement



National Data Curation Team, in Partnership with TRUST and BII to Serve Ecosystem Data Curation and Standards Needs

The primary goal of the National Data Curation Team (NDCT) is to support and uplift the local data ecosystem in terms of data quality and interoperability.

Bioinformaticians, Data Scientists, Software Engineers

1. **Identify, develop and promulgate appropriate data standards**, including Minimum Information Standards for experimental data

2. **Create and maintain community resources for data management** (e.g. data cleaning pipelines, curation tools etc)

3. **Train and uplift ecosystem in data management**

4. **Develop a network of "Communities of Practice"** for different data types to support above activities

<u>Data types under consideration</u>

clinical data | genomics | transcriptomics | single-cell and spatial omics | proteomics | metabolomics

BDH Costing Model

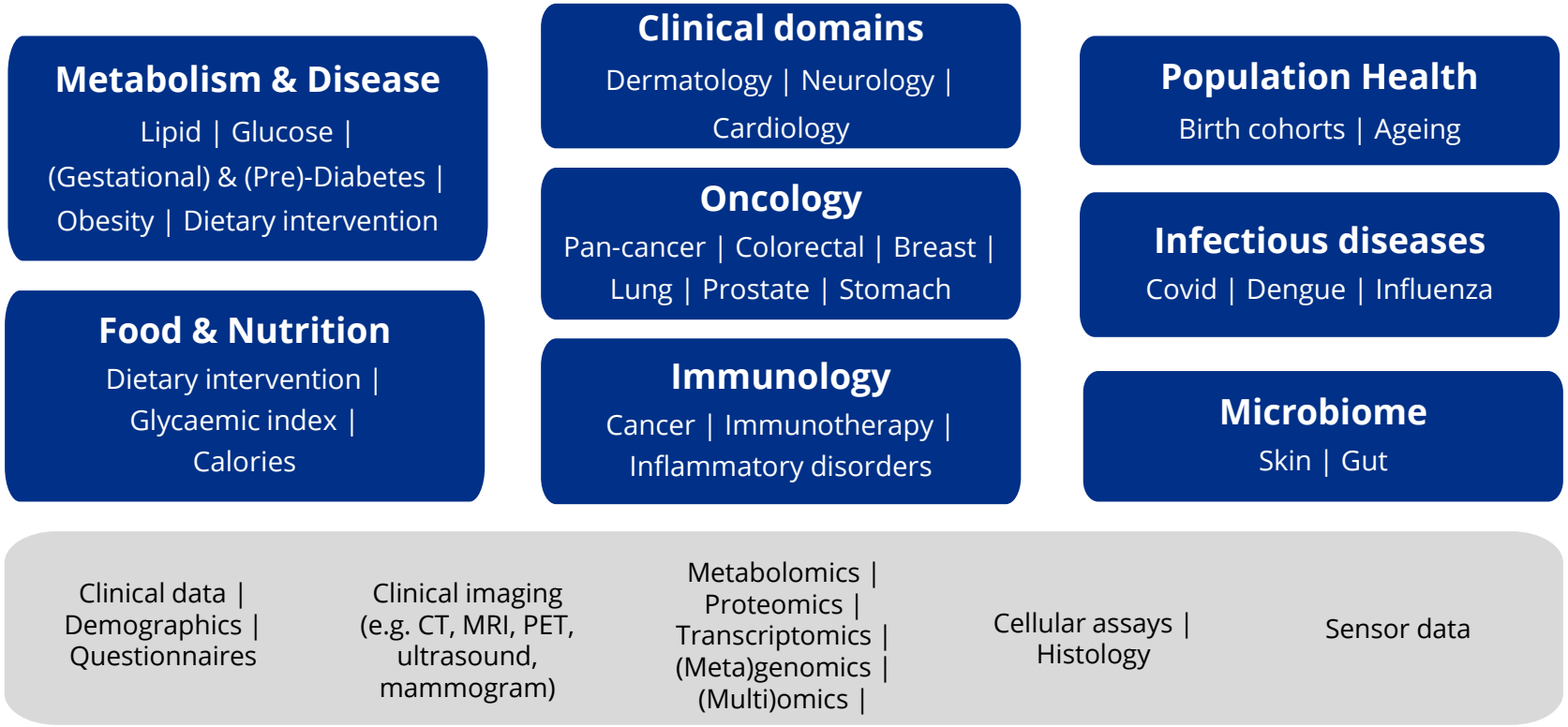
	Data Owners (Repository Costs)	Secondary Users (TRE costs)
Storage	• Term storage (fee-for-service) • Lifecycle storage (1X fee for lifetime storage) • 50% reduced fee for data sharing in catalog • Charged to BDH core funds (for strategic cohorts), NMRC, or other grants	• A*STAR users: Project storage and compute costs billed according to A*CRC / A*CC rate card • Other users aligned with TRUST charging framework:
Compute	• Data curation compute and NDCT services charged as part of SAs and RCAs • AI data curation compute and licensing costs passed through with administrative charge • Fees borne by grants, BDH core funds (for strategic cohorts), or NCO grantees	
Services		• Clinical or bioinformatics expertise availed to research or industry investigators to support analysis via SA or RCA

Monthly Subscription Rates		Brief Description	Computing Configurations	
Base Package	Cost per user per month		Workspace Package – preconfigured with 160 hours of usage per user per month	
Workspace with Small SageMaker	$170	Ideal for small datasets and basic data exploration tasks, such as simple data cleaning and visualization. Not suitable for large-scale training or computationally intensive tasks.	Workspace: 2 vCPU, 4GB Memory, 50GB user storage	Small SageMaker: ml.t3.medium c/o: 2 vCPU, 4GB Memory, 20GB storage
Workspace with Medium SageMaker	$190	Suitable for moderate-sized datasets and more complex data processing tasks, including feature engineering and model training on smaller models.		Medium SageMaker: ml.m5.large c/o: 2 vCPU, 8GB Memory, 20GB storage
Workspace with Large SageMaker	$220	Well-suited for large datasets and computationally intensive tasks, such as training deep neural networks or running large-scale simulations.		Large SageMaker: ml.t3.xlarge c/o: 4 vCPU, 16GB Memory, 20GB storage
Workspace with X-Large Sagemaker	$420	This is a more powerful general-purpose instance with significant compute and memory resources. Suggested to use in tasks like handling large-scale data processing jobs or supporting multiple concurrent data analysis workloads.		XL Sagemaker ml.m5.4xlarge c/o: 16 vCPU, 64GB Memory, 20GB storage
Workspace with GPU	$390	Designed for tasks that require significant parallel processing, such as training deep learning models, computer vision, and natural processing. Ideal for applications that can benefit from GPU acceleration.		GPU: ml.g4dn.xlarge c/o: 4 vCPU, 16GB Memory, 20GB storage, 1 GPU

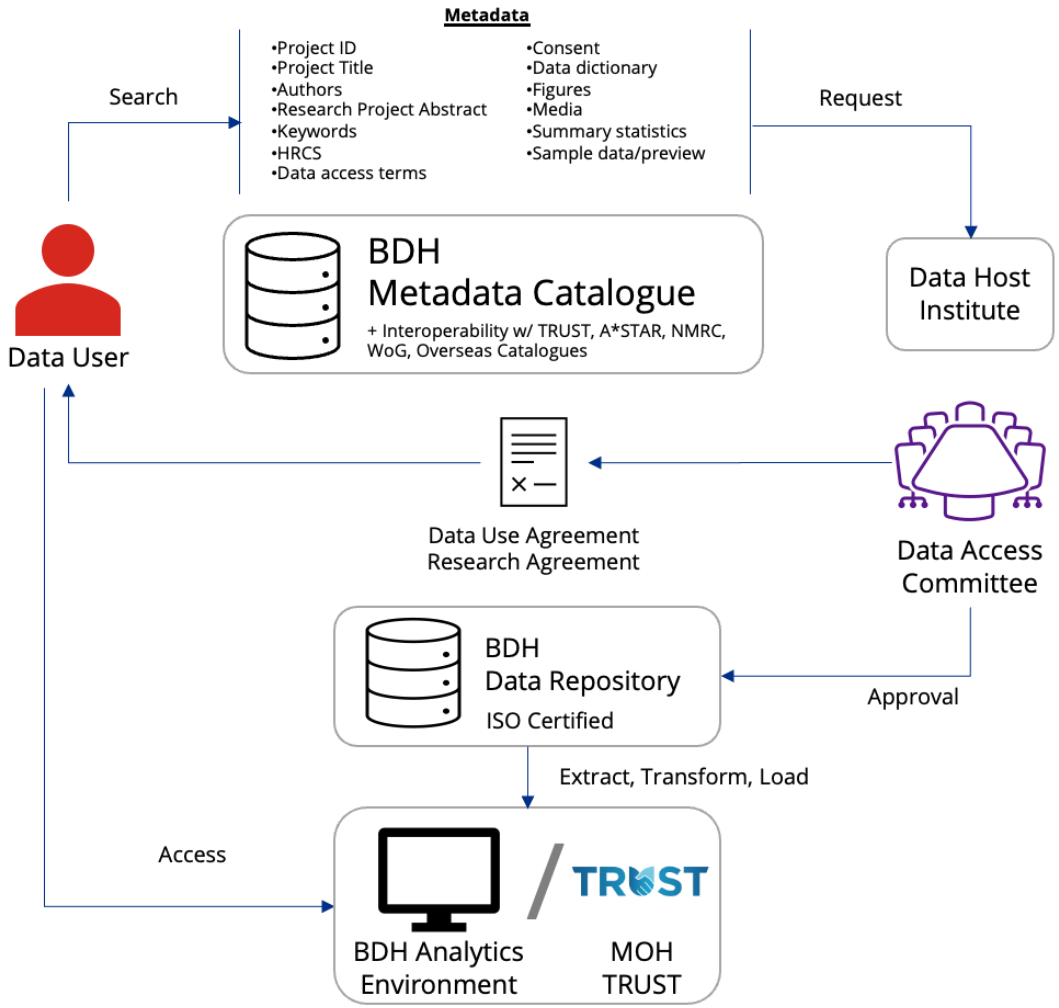
Trust Charging Framework

(Subject to revision based on BMSIPO Study)

Examples of Biomedical Datasets Accessible with DAC Approval



Recap: BDH helps makes Singapore research data Findable, Accessible, Interoperable and Reusable (F.A.I.R.)



Use Cases

Discovery

Search for relevant data and quickly assess its utility

Governance

BDH tech infra ensures data owners retain control over access and use

Analysis

Data is accessed and analysed in a separate and secure environment

Capabilities

Catalogue

User searchable metadata catalogue *(Coming Soon)*

Repository

ISO certified secure multi-modal data repository

Analytical Node

Trusted research environment (TRE) matched to compute and linkable to clinical data[1]

National Data Curation Team

Coordinates data standards working groups and a federated resource pool of data curators

Project Management

Experienced bioinformatians to help facilitate access and use of data for research and innovation

Data Types and Data Sets

- Deidentified research data
- Multiple clinical domains (Cardiology, oncology initially)
- Multi-modal data: Tabular, free-text, imaging, omics, ECG

(1) Clinical data linkage via MOH TRUST

Agency for Science, Technology and Research
SINGAPORE

THANK YOU

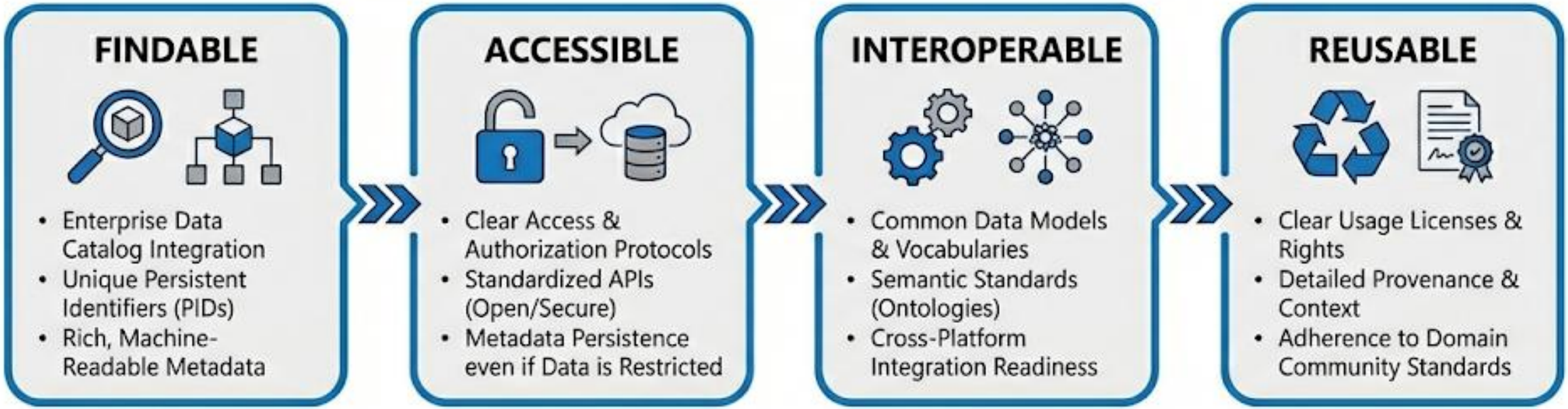
ASTARSG ASTAR-SG

David Vu
Program Director, BDH
david_vu@bdh.sg
+65 9155 1492

Data Repository Classifications

	Biospecimen-Linked Repository	Digital Data Repository	Study & Registry Repository
Physical specimens	✓	✗	✗
Digital datasets	✓	✓	Partially (mostly summaries)
Metadata	✓ (extensive)	✓	✓ (high-level)
Core Characteristics	• Stores physical specimens + linked digital data • Requires cold storage, inventory systems, biosafety protocols	• Stores large-scale digital biomedical data only • Typically, federated or centralised architectures • Designed for secure access, compute, and secondary use and reuse at scale • Strong emphasis on interoperability and standards	• Focused on studies, not individuals • Often high-level summary data • Limited or no linkage to underlying datasets
Data Types	• Clinical-linked data	• Genomic / omics data • Medical imaging (radiology, pathology) • Clinical datasets (individual-level, de-identified) • Sensor / device / longitudinal health data • Rich metadata	• Study protocols • Summary statistics • Trial registration details • Aggregate outcomes • Incidence/surveillance data
Reusability	• Highest	• High (analysis-ready)	• Moderate (discovery oriented, not analysis-ready)
Strengths	• Enables new assays in the future • Supports precision medicine, translational research, and biomarker discovery • Highest long-term scientific value	• Easier national and global sharing • Can handle large datasets • Supports large scale analysis	• High visibility • Enables discovery and research landscape mapping • Lighter governance burden
Limitations	• Expensive to maintain • Complex governance and consent • Operationally complex – shipping, sample safety & preservation challenges	• Scientific value highly dependent on metadata quality • Requires strong data governance and standardisation upfront	• No granular data • Cannot support deep analytics • Often not linkable to datasets
Examples	EU: *BBMRI.it* (Italy) ASIA: Korea Biobank Network (Korea)	US: The Cancer Imaging Archive (TCIA), Genomic Data Commons (GDC) EU: Health-RI (Netherlands), Health Data Hub (France) ASIA: Tohoku Medical Megabank (Japan)	US: ClinicalTrials.gov EU: German Center for Cancer Registry Data (ZfKD) (Germany) ASIA: Korean Genome and Epidemiology Study (KOGES) (Korea)

FAIR Data Principles



5-Safes framework for data governance ensures secure data sharing