The *BioMed DAR* (*Data Architecture & Repository*) Programme for BMRC clinical research data management

Wong Wing-Cheong
Head & Principal Investigator,
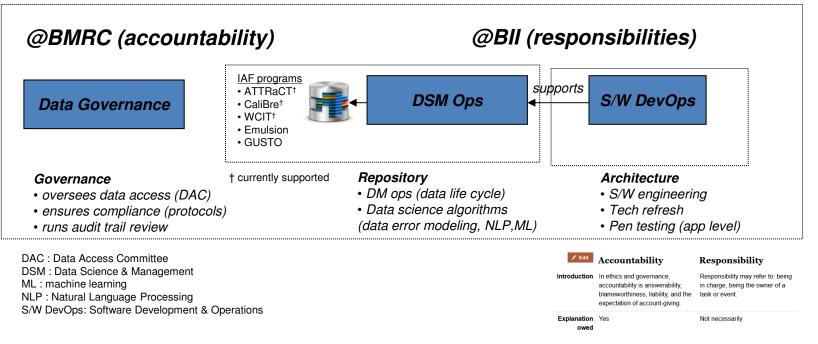BioMed DAR, BMRC

April 2021

BioMed DAR: Mission and Structure

Mission statement :

To operationalize a SSSO (Standard Systems Support Office) for clinical research data management to support strategic (past, current, future) A*STAR BMRC programmes
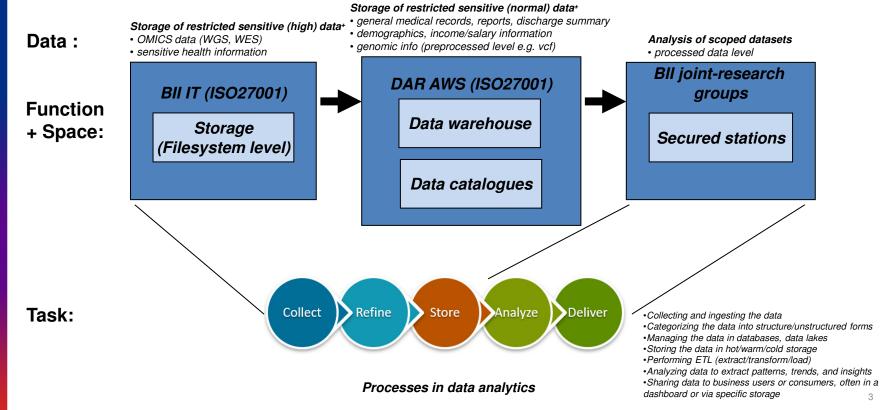
Structure :

@BMRC (accountability)   *@BII (responsibilities)*

Data Governance	**DSM Ops**	**S/W DevOps**

IAF programs
- ATTRaCT†
- CaliBre†
- WCIT†
- Emulsion
- GUSTO

supports

Governance
- oversees data access (DAC)
- ensures compliance (protocols)
- runs audit trail review

† currently supported

Repository
- DM ops (data life cycle)
- Data science algorithms (data error modeling, NLP,ML)

Architecture
- S/W engineering
- Tech refresh
- Pen testing (app level)

DAC : Data Access Committee
DSM : Data Science & Management
ML : machine learning
NLP : Natural Language Processing
S/W DevOps: Software Development & Operations

	Accountability	Responsibility
Introduction	In ethics and governance, accountability is answerability, blameworthiness, liability, and the expectation of account-giving.	Responsibility may refer to: being in charge, being the owner of a task or event.
Explanation owed	Yes	Not necessarily

✏ Edit

CREATING GROWTH, ENHANCING LIVES

BioMed DAR: BII integration

Technical integration of BII-DAR (FY2021) :

Data :

Storage of restricted sensitive (high) data+
- *OMICS data (WGS, WES)*
- *sensitive health information*

Storage of restricted sensitive (normal) data+
- *general medical records, reports, discharge summary*
- *demographics, income/salary information*
- *genomic info (preprocessed level e.g. vcf)*

Analysis of scoped datasets
- *processed data level*

Function + Space:

BII IT (ISO27001)

Storage (Filesystem level)

DAR AWS (ISO27001)

Data warehouse

Data catalogues

BII joint-research groups

Secured stations

Task:

Collect → Refine → Store → Analyze → Deliver

- *Collecting and ingesting the data*
- *Categorizing the data into structure/unstructured forms*
- *Managing the data in databases, data lakes*
- *Storing the data in hot/warm/cold storage*
- *Performing ETL (extract/transform/load)*
- *Analyzing data to extract patterns, trends, and insights*
- *Sharing data to business users or consumers, often in a dashboard or via specific storage*

Processes in data analytics

+ Data classification level based on MOH HIMS-DM document Table 5

CREATING GROWTH, ENHANCING LIVES

3

BioMed DAR: Technical Stack

Data Request

DRAS†
(in-house s/w)

Public-facing network

Data warehouse

Master DB

TIMS
(in-house s/w)

ETL (Extract, Transform, Load)

ETL DB

Data lake (unstructured)

Data architects + data engineers

Master data catalogue

DashStats (in-house s/w)

VPN

ISO27001 compliance network (main/syn/cat)

Scoped DB

DataServ
(in-house s/w)

Sandbox environment

docker

Data Catalogue

Clinicians & data scientists

VPN

1. Data oversight and KPI reporting to stakeholders
2. Data catalogue for data scientists

ISO27001 compliance network (usage)

Technical stack aims to support :
- Data life cycle activities (maintenance, synthesis, cataloguing, usage, archival, purging)
- Data governance and oversight (approvals, risk management, data security/protection)
- Digital audit review (who, what, when & why)

† to be built

CREATING GROWTH, ENHANCING LIVES

4

BioMed DAR: Datasets & Data life-cycle

Datasets and data life-cycle stages :

Dataset	Maintenance	Synthesis	Cataloguing	Usage	Archival	Purging
ATTRaCT	☑	☑	☑	☑		
CaliBre	☑	☑	☑			
WCIT	☑	☑	☑			

Incoming: GUSTO, Emulsion

About Data life cycle:
The sequence of stages that a particular unit of data goes through from its initial generation or capture to its eventual archival and/or deletion at the end of its useful life.

Capabilities associated to data life-cycle ops:
1. Maintenance (cleansing, integration, movement, governance)
2. Synthesis (data creation from existing data e.g FASTQ to vcf)
3. Cataloguing (discovery, describe, organize)
4. Usage (data access enabling and sandboxing)

Capabilities supporting data life-cycle ops:
1. s/w engineering (data warehouse, catalogue and request/approval systems)
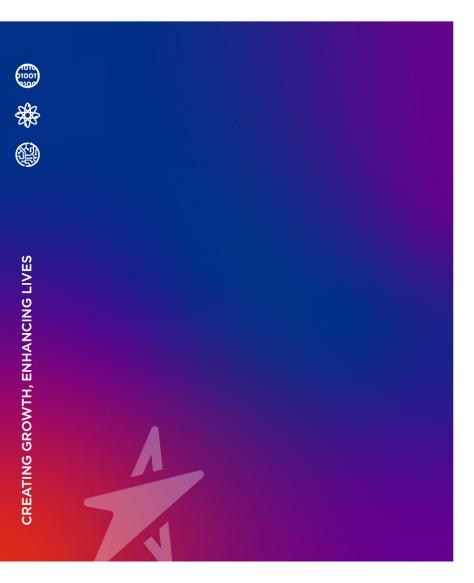
CREATING GROWTH, ENHANCING LIVES

Biomedical DAR: User case

Data Governance & Data access enabling:

Project proposal
(Academics, Industries)

Review & evaluate

Institutes
(SBIC,BII,CVRI,IMCB,NHCS)

DAC

Health clusters
(Singhealth, NUHS, NHG)

JSC

Higher level approval

ATTRAcT DataStore

Once approved, data extraction triggered based on (i) data scope (ii) data usage (iii) project duration

BII/DAR

After the appropriate data sets are identified, the schema will then be defined for the ETL process to create the scoped database for the project

A copy of the data base will be cloned and made available on a secured station in BII.

Analytics

BII premises

Analytics

Cloned DB — **Secured workstation** — *Data scientists*

Dashboard

BII VPN

Stakeholders

Analytics (aggregated level)

Data scientists

Master DB

User-defined schema

ETL

Scoped DB

API server

datapoints

aggregated data (via API)

Data store

Delivery

Data store comprises of both data warehouse (structured data) and data lake (file-based data).

Structured data refers to annotated clinical information and measures. File-based data refers to pre-processed OMICs data (e.g. RNA count, VCFs)

Data is accessible via 2 options:
1.BII premise – all data points can be analyzed on a secured station within premises but data cannot be copied.

2.BII VPN – aggregated data can be retrieved via an API library invoked by the user's programming scripts situated elsewhere.

Dashboard serves as a data catalogue to provide an overview description & content of the scoped datasets.

ETL: Extract/Transform/Load
DB: Database
API: Application Programming Interface

ISO27001 compliance network

JSC : Joint Steering committee
DAC : Data Access Committee (PIs of awarded IAFs)

CREATING GROWTH, ENHANCING LIVES

THANK YOU

www.a-star.edu.sg