AHRI AFRICA HEALTH RESEARCH INSTITUTE

Reference Death Archive (RDA)

Pilot Demonstration

RDA Development Team

Samuel Clark
Kobus Herbst
Yue Chu
Doris Ma Fat
David Plotner
Norman Goco
Philippe Boucher
Jean-Francoise Saint-Pierre
Joven Larin
Jason Thomas
Mona Sharon

Aims

1. Create a secure infrastructure to accept, clean, document, store, and manipulate, reference deaths with VA, MITS, and reference cause
2. Create a publicly-accessible repository of de-identified data products from reference deaths

Together (1) and (2) form the Reference Death Archive (RDA)

3. Create a set of procedures, roles, and responsibilities to operate the RDA
4. Populate the RDA with reference deaths from the CHAMPS, COMSA, and MITS Alliance projects and the SVO mortality surveillance system in the city of Sao Paulo, Brazil
5. Host the RDA at the WHO with WHO URL and WHO-linked DOIs for data

Concepts

- Fully general, metadata-driven data model
  - underlying data model and supporting code do not change
  - all customizing of the datastore and/or code behavior accomplished using metadata
  - harder to build, much easier to maintain, much more flexible, and likely to survive longer

- Store all related metadata and paradata with raw data
  - full history, semantics, and provenance of each datum retained and propagated through creation of derivative datasets
  - creates browsable and searchable metadata stores for data exploration

- Separate security from data management
  - make security an encapsulating service consistent with WHO policies
  - optimize data management for speed within a secure environment

Concepts

- Confidentiality
  - adhere to WHO policies governing data confidentiality
  - plan to eventually handle of most sensitive individual information
  - individual-level data can come into archive but never leave
  - within secure WHO container, create ‘trusted research’ environment for free manipulation of reference deaths
  - create entirely separate repository for accessible data products, and a actively controlled vetting system to transfer not individually identifiable data products to the repository and possibly release them for public access

- **Support for Reproducible Research and Full Data Provenance**
  - full metadata reuse and code used to prepare dataset is archived with dataset
  - final datasets required to include comprehensive documentation
  - only one route for data egress from archive - same for dataset producers and other users

Implementation Principles

- RDA implemented within a virtual machine (VM)
  - VM host provides security consistent with WHO policies
  - VM runs software for RDA trusted researcher environment and archive

- Trusted Research Environment
  - secure access control
  - verifiable researchers
  - stores primary, ‘raw’ data
  - coding sandbox with access to raw data using R, Python, or Julia
  - no direct data download
  - automated data export to repository with request for review for confidentiality requirements
  - successfully reviewed data released via repository either publicly or to verified users

- Data Release Repository
  - full catalog of browsable, searchable metadata describing the raw data
  - fully documented release data that pass confidentiality requirements with DOI
  - data use agreement management

Implementation Principles

- **Non-proprietary, widely-used toolset, all code managed using GitHUB**

- Underlying SQL Schema does not change
  - entity-attribute-value (EAV) inspired design
  - highly abstract
  - fully normalized
  - extensive use of metadata

- Metadata
  - source information
    - institution
    - study protocol(s)
    - ethics approvals
    - data collection instruments
  - data dictionary
    - variable definitions
    - standardised vocabularies

Implementation Principles

Speed and implementation efficiency

- Data store: **SQLite**
  - primary use is analysis, no need for advanced transactional support (e.g. SQL Server, PostgreSQL)
  - user access managed at application level, no need for sophisticated user access management (also e.g. SQL Server, PostgreSQL)
  - structured data, no need for NoSQL database

- Supporting code: **Julia**
  - used for RDA management - highly performant for back office data management
  - additional support for R and Python

Technology Overview

JupyterHUB

- Supports RDA data store and core operations under admin accounts
  - create, maintain database
  - ingest, clean, document raw data
  - export approved data products to archive
- programming environments for all accounts: *R, Python, Julia*
- example notebooks to illustrate common use cases
- secure access based on ORCIDs
- *no upload/download for regular users*
- identify data products for approval and export to repository

Technology Overview

NADA Repository

- open source data repository system created/maintained by *World Bank*
- hosts data catalog and approved, accessible data products
- browsable, searchable data catalog for secondary data, but not the raw data themselves
- browsable metadata
- data use agreement management
- download management
- citation management incorporating DOIs

High-level Design

Data producer

- CHAMPS
- COMSA – Mozambique
- HEALSL

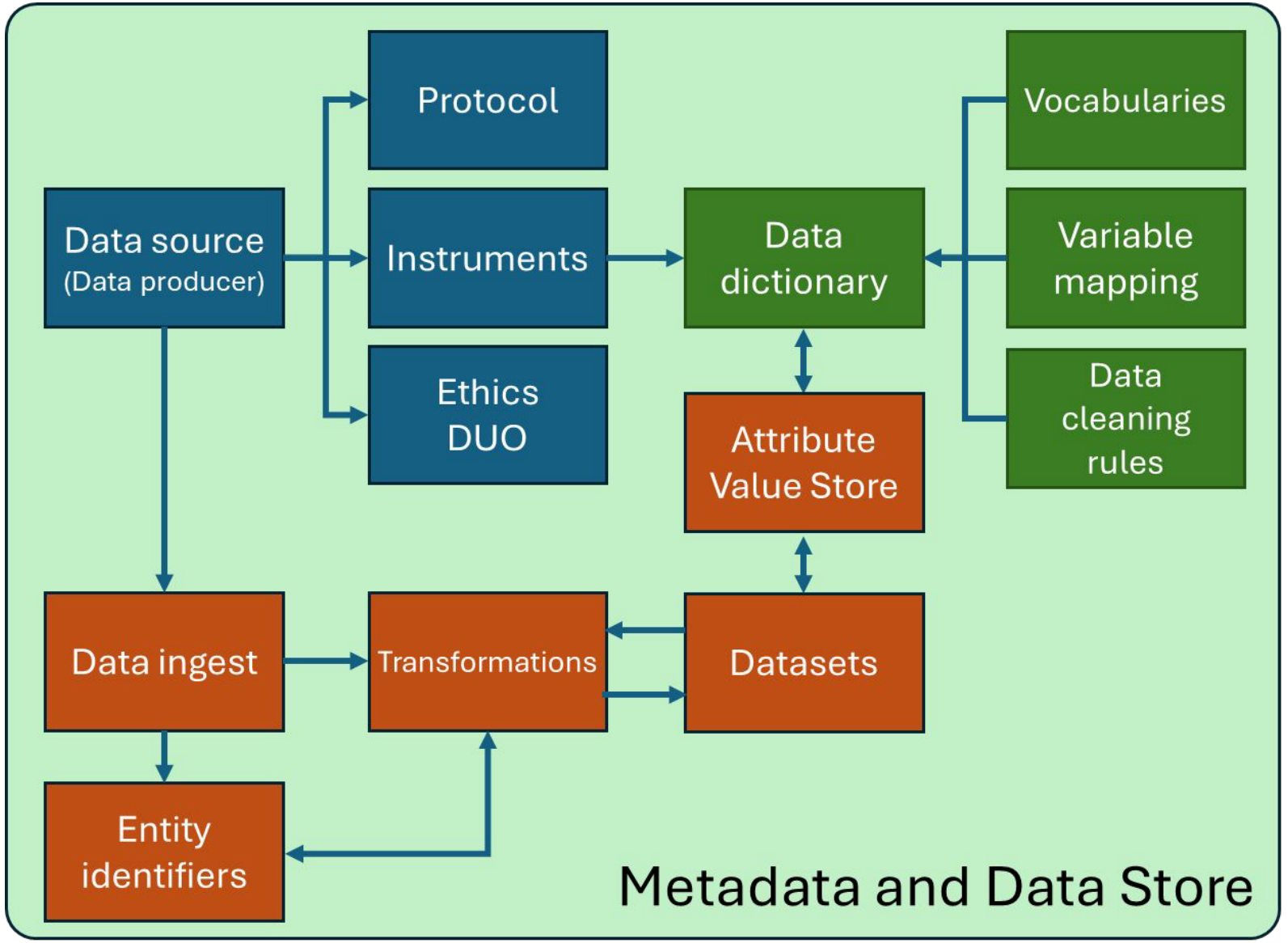
Reference Death Archive

- Documentation
- Data linkage / synthesis
- Harmonization

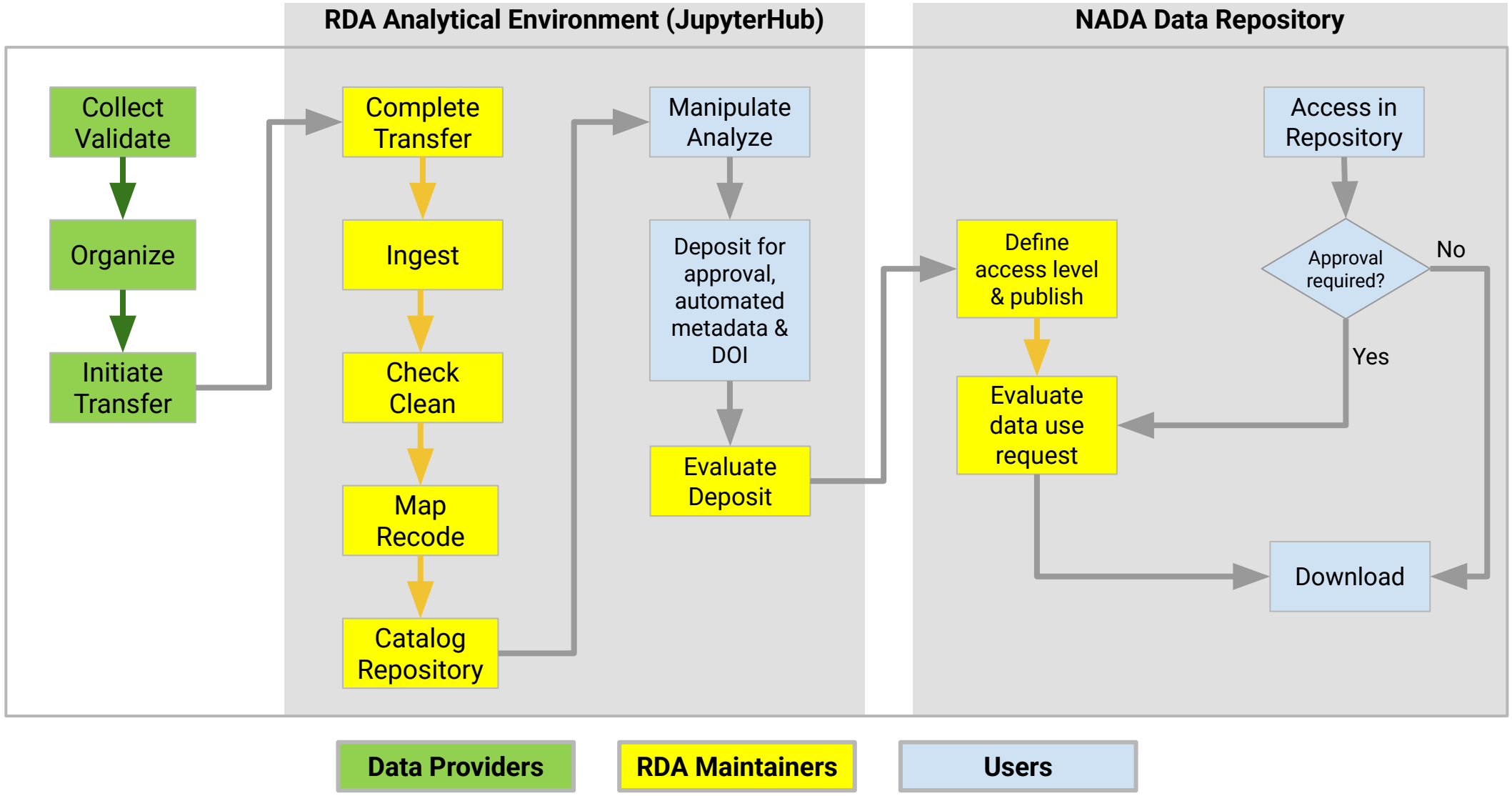
Data User

- JupyterHub
- National Data Archive (NADA)

SQL Schema - Data Model



Data Flow


RDA Analytical Environment (JupyterHub)
NADA Data Repository
Collect Validate
Organize
Initiate Transfer
Complete Transfer
Ingest
Check Clean
Map Recode
Catalog Repository
Manipulate Analyze
Deposit for approval, automated metadata & DOI
Evaluate Deposit
Define access level & publish
Evaluate data use request
Access in Repository
Approval required?
No
Yes
Download
Data Providers
RDA Maintainers
Users

Acknowledgements

Investigators

- OSU and overall PI: Samuel J. Clark
- WHO PI: Doris Ma Phat
- AHRI PI: Kobus Herbst
- RTI PI: Norman Goco
- USP PI: Luiz Fernando Ferraz da Silva (Burns)

Key collaborating institutions

- The Ohio State University (OSU)
- WHO, Department of Data and Analytics
- Africa Health Research Institute (AHRI)
- Research Triangle Institute (RTI)
- University of Sao Paulo (USP), Department of Pathology

Launch data contributors

- Research Triangle Institute (RTI) and Minimally Invasive Tissue Sample (MITS) Alliance
- CHAMPS Project
- COMSA Projects - Mozambique and Sierra Leone
- City of Sao Paulo, Brazil mortality surveillance system

Funder: Bill and Melinda Gates Foundation

Discussion