DS-I Africa Open Data Science Platform

Nicky Mulder
University of Cape Town
NIH DSI-Africa

• Harnessing Data Science for Health Discovery and Innovation in Africa
  • **Research Hubs:** Advance and demonstrate feasibility of data science research and innovation to improve health in Africa
  • **Training:** Increase capacity for data science research in Africa
  • **ELSI Research:** Explore Ethical, Legal, and Social Implications of data science research from an African perspective and contribute to policy discussion on the continent
  • **Open Data Science Platform & Coordination Center:** Facilitate the development of a trans-African network of data scientists

Image and text extracted from: [https://commonfund.nih.gov/africadata](https://commonfund.nih.gov/africadata)
Funded project titles (1)

Research hubs

• Role of Data Streams In Informing Infection Dynamics in Africa- INFORM Africa
• UZIMA-DS: UtiliZing health Information for Meaningful impact in East Africa through Data Science
• Developing data science solutions to mitigate the health impacts of climate change in Africa: the HE2AT Center
• Harnessing Data Science to Promote Equity in Injury and Surgery for Africa
• Combatting AntiMicrobial Resistance in Africa Using Data Science (CAMRA)
• MADIVA (Multimorbidity in Africa: Digital innovation, visualisation and application)
• MUST Data Science Research Hub (MUDSReH)
Funded project titles (2)

ELSI:

- Research for Ethical Data Science in Southern Africa (REDSSA)
- Bridging Gaps in the ELSI of Data Science Health Research in Nigeria (BridgELSI)
- DS-I Africa - LAW
- Public Understanding of Big data in Genomics Medicine in Africa (PUBGEM-Africa)

7 Training grants on different aspects of data science & health

Open Data Science Platform & Coordinating Centre
Open Data Science Platform (ODSP) aim

• To develop an African Open Data Science Gateway, including the eLwazi platform and associated resources, to support the Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa) consortium and beyond. eLwazi, will be a flexible, scalable Open Data Science Platform enabling the implementation of data science for health, that is relevant to the African context.

Ulwazi is the Xhosa word meaning “knowledge” or “information”, and Olwazi means big rock in Luganda

eLwazi
AIM 1: Build Infrastructure

- Cloud computing
- Workflows
- eLwazi Data Science Gateway
- Anvil
- Software stack
- iRODS
- Local EGA
- Users to store and analyze data
- Data Discovery Service
- Platform API
- Security and access (AAI)
- Cloud and local computing hosts
Aim 2. Access to data and resources

- Collect data types
- Reference datasets
- Metadata stds and curation, FAIR
- Data harmonization
- Cohorts atlas, catalogue
- Linking datasets
- Data Portals
- Other resources: SOPs, protocols, training materials
eLwazi Data Science Gateway

AIM 2: Data, metadata, catalogues

Metadata catalogue

Data portals
- AGVD
- APDB
- AMP
- Beacon

Public

Transform data into usable form
DB vs KB

Metadata catalogue

Software stack
- Software stack
- Data
- Metadata
- Ref data
- User workspaces

Anvil

Training content

Pathogens Protocols

Transform data into usable form

Workflows

Cloud computing

Users to store and analyze data

DB vs KB

Ref data

Metadata
catalogue

Data portals

 AIM 2: Data, metadata, catalogues

AGVD
APDB
AMP
Beacon
Aim 3. Tools and workflows

- Identify tools
- Installation on infrastructure
- Workflows
- Interface with Cloud providers
- Tool and workflow registries
- Galaxy
- Beginner and advanced access to tools
AIM 3: Tools & workflows

User workspaces
Controlled access
Users to store and analyze data

Metadata catalogue
Data portals: AGVD, APDB, AMP, Beacon

Transform data into usable form
DB vs KB

Software stack

Public

Transform data

Ref data
Meta data

Data

User workspaces

Data Discovery Service
Platform API

Transform data into usable form

DB vs KB

Anvil

Containerize workflows
Add to registry
WES

Interface for access to tools and workflows

Tool registry
Tool suite:
- ML
- Genomics
- Other

User tools: build cohort, build dataset, Jupyter notebooks

Pathogens Protocols

Training content

Data portals

elwazi Data Science Gateway
DS-I Africa ODSP hosting sites

Cloud & HPC Environments
- AWS, Azure, and Local

DRS for data access

WES for compute

TRS for workflow sharing

Workspaces for:
- Compute on data in various locations
- Interactively analyze via Notebooks
- Sharing and Collaborating
Aim 4. User support

• User support type
  • User group meetings
  • Use case support
  • Clinical data support

• Helpdesk

• Training
  • Curriculum standards
  • Professional training
  • Coordinated with training programs
eLwazi Data Science Gateway

AIM 4: User support

- Metadata catalogue
- Data portals: AGVD, APDB, AMP, Beacon
- User workspaces
- Cloud computing

Workflow:
- Transform data into usable form
- DB vs KB
- Ref data
- Meta data
- Software stack
- User workspaces
- Data Discovery Service
- Platform API
- ELSI
- AAI
- Helpdesk

Public:
- Transform data into usable form
- DB vs KB
- Meta data catalogue
- Data portals
- User tools: build cohort, build dataset, Jupyter notebooks

Research hub:
- Access for novice and advanced users

Data standardization & harmonization:
- FAIR
- REDCap support
- PRDB

User tools:
- Build cohort, build dataset, Jupyter notebooks

Training content:
- Training Grants
- REDCap support
- PRDB

Pathogens Protocols

iRODS Local EGA

Training content:
- Grants
- Pathogens Protocols
Aim 5. Monitoring, outreach, sustainability

• Monitor platform usage and impact
• Dissemination
• Develop cost model and sustainability plan
• Engagement with external stakeholders
  • NRENS
  • Cloud providers
  • Industry
Overview of the ODSP

- Biomedical Data
- Dictionary Harmonization
- Metadata Models

- eLwazi
  - Dictionary Harmonization
  - eLwazi Data Catalogue
  - Data Repo

- eLwazi User Workspaces
  - Search data
  - Retrieve datasets
  - Choose tools
  - Select compute environment

- ODSP Gateway
  - Search data
  - Retrieve datasets
  - Choose tools
  - Select compute environment

- Expandable to other data, tools, applications
- Public data portals
- Training materials, SOPs
- Documentation Helpdesk

- eLwazi Data Catalogue
- ODSP Gateway
- BAWS
- Azure
- Dockstore
- GA4GH

- AW
- DR
- TR
- WE
- Local

- Results

- eLwazi Training materials, SOPs

- Public data portals
  - Search data
  - Retrieve datasets
  - Choose tools
  - Select compute environment

- Expandable to other data, tools, applications
Acknowledgements

**Coordinating Centre PI:** Dr Michelle Skelton

**ODSP Partners**

- University of Cape Town
- University of the Witwatersrand
- University of the Western Cape (Ilifu)
- University of Kwazulu-Natal -> Stellenbosch
- University of Mauritius
- Uganda Virus Research Institute
- University of Khartoum
- USTTB, University of Bamako
- Broad Institute
- EMBL European Bioinformatics Institute
- University of California, Santa Cruz
- University of Chicago

**Funding:** DS-I Africa program through the NIH Common Fund