## 

HE²AT Center Data Management Plan

Document version number 3.0

Version Date: /2024

Developed by \_\_Christopher Jack\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name Date Signature

Reviewed by \_\_NIH Science officer tbd\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name Date Signature

Approved by \_\_Matthew Chersich\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name Date Signature

# 

# Table of contents

[Table of contents 2](#_Toc173777770)

[1. Scope 4](#_Toc173777772)

[2. Purpose 4](#_Toc173777773)

[3. Abbreviations 5](#_Toc173777774)

[4. Definitions 6](#_Toc173777775)

[5. Background 6](#_Toc173777776)

[6. Data categories 8](#_Toc173777777)

[7. Data management workflow 12](#_Toc173777778)

[7.1. Setting Up the Data Transfer Agreement (DTA) 12](#_Toc173777779)

[7.2. Ethics Approval for New Databases 12](#_Toc173777780)

[7.3. Data encryption and transfer 13](#_Toc173777781)

[7.4. Data storage and encryption 14](#_Toc173777782)

[7.5. Data indexing 14](#_Toc173777783)

[7.6. De-identification 15](#_Toc173777784)

[7.7. Codebook remapping and harmonisation (Peter – Renaming this section) 15](#_Toc173777785)

[7.8. Health Data Harmonisation 15](#_Toc173777809)

[7.9. Data integration and analysis (Revise – integrate w/above section; focus on the development of integrated relational database) 16](#_Toc173777810)

[7.10. Data analysis platforms 17](#_Toc173777811)

[8. Data Management, Documentation and Curation 17](#_Toc173777812)

[9. POPIA compliance and protection of personal information 18](#_Toc173777813)

[10. De-identification 20](#_Toc173777814)

[10.1. Geographic Aggregation: 20](#_Toc173777815)

[10.2. Location Jittering: 20](#_Toc173777816)

[11. Data encryption 22](#_Toc173777817)

[12. Storage isolation 22](#_Toc173777818)

[13. Network firewall and Virtual Private Network 22](#_Toc173777819)

[14. Local authentication and authorisation 23](#_Toc173777820)

[15. Data Sharing and Open Access 23](#_Toc173777821)

[16. Procedure for making data available to qualified individuals 24](#_Toc173777822)

[17. Governance of data sharing 26](#_Toc173777823)

[18. Data Access Committee 26](#_Toc173777824)

[19. Data Retention 28](#_Toc173777825)

[20. Roles and Responsibilities 28](#_Toc173777826)

[21. Assessment and revision 30](#_Toc173777827)

[22. References 30](#_Toc173777828)

[Annex 1: Key data sources 1](#_Toc173777829)

[Annex 2: Personal information processing agreement 7](#_Toc173777830)

[Annex 3: Ethics notification letter 8](#_Toc173777831)

## 

# 

# Scope

This Data Management Plan (DMP) applies to all data acquired and produced as part of the activities of the HE²AT Center, including pilot projects. The scope encompasses various stakeholders, each with distinct roles and responsibilities as outlined below:

1. **Data Providers**: These are the legal entities or organizations responsible for authorizing access to their datasets. They ensure that data sharing complies with legal and ethical standards.
2. **DSI Partners**: Partners within the NIH Data Science Initiative (DS-I) Africa contribute to and utilize the shared data resources. They collaborate on data management practices and ensure alignment with the broader DS-I Africa objectives.
3. **Consortium Research Groups**: These groups, which are part of the HE²AT Center consortium, conduct research using the shared datasets. They adhere to the DMP guidelines for data acquisition, transfer, processing, storage, and access.
4. **El Wazi**: As a data management platform, El Wazi facilitates the indexing, storage, and sharing of data within the consortium. It ensures that data is discoverable, accessible, interoperable, and reusable (FAIR principles).
5. **Training Hubs**: These hubs provide training and capacity-building activities for researchers. They utilize the shared data for educational purposes and contribute to the development of data management skills within the consortium.
6. **Researchers Looking to Acquire Data**: External researchers seeking access to the HE²AT Center's datasets must follow the data access procedures outlined in this DMP. They are required to comply with data sharing agreements, ethical guidelines, and any conditions set by the Data Access Committee (DAC).

# Purpose

**Purpose**

The purpose of the Data Management Plan (DMP) is to establish comprehensive procedures and standards for the acquisition, transfer, processing, storage, and access to data within the HE²AT Center program. This plan aims to ensure the integrity, security, and availability of data while facilitating effective data sharing and collaboration among all stakeholders.

Specifically, the DMP seeks to:

1. **Standardize Data Management Practices**: Define uniform procedures and standards to streamline data handling processes across the consortium, ensuring consistency and reliability.
2. **Enhance Data Security and Privacy**: Implement robust measures to protect sensitive and personally identifiable data, complying with relevant legal and ethical standards.
3. **Facilitate Data Sharing and Access**: Support the sharing of data within the consortium and with external researchers through clear guidelines and Data Sharing Agreements, promoting transparency and collaboration.
4. **Support Data-Driven Research**: Provide a framework that enables researchers to efficiently access and utilize data for innovative research, contributing to the HE²AT Center's mission of advancing heat-health knowledge and interventions.
5. **Ensure Compliance with Ethical Guidelines**: Maintain adherence to ethical standards and guidelines, including obtaining necessary approvals and consents for data use.

While this DMP outlines the procedures for data access by external parties, detailed data sharing outside the consortium is further supported by associated Data Sharing Agreements, ensuring that all data management activities align with the HE²AT Center's objectives and ethical commitments.



# Background

The HEat and HEalth Africa Transdisciplinary Center, HE²AT Center, is a U54 Cooperation agreement with the NIH (2021-2026). The HE²AT Center aspires to become a Center of Excellence in heat-health research, capacity building and engagement, using population health science and applying data science methodologies to improve the health of populations in Africa and beyond. The goal of the HE²AT Center is to advance the development of new health knowledge and human capacities through reusing existing data to generate and then disseminate heat-health knowledge and innovations.

**RP1 description**

Research project 1 is an Individual participant data (IPD) meta-analysis to assess the size and shape/nature of associations between exposure to high ambient temperatures and selected maternal and child conditions within the first two years of life. Such techniques have not yet been employed in climate change and health. They can overcome many limitations of traditional analyses of individual datasets and biases in classic systematic review methodology. The project will systematically identify potentially eligible African cohort studies or trials through systematic mapping reviews. Data will be harmonised through re-coding raw individual participant data into a standard set of variables. Subsequently, all the individual participant's data from each eligible study will be pooled. Analyses that include a range of traditional statistical and novel machine-learning approaches will quantify associations between exposure to high temperatures and adverse maternal and neonatal outcomes. The study may provide robust, definitive evidence on the impacts of heat on maternal and child health and allow for estimation of the burden of rises in temperatures and other climate change manifestations on maternal and neonatal health[1].

**RP2 description**

Rapid urban growth, significant levels of informality and increasingly stretched health services, intersecting with observed past and projected future temperature increases, have resulted in a critical emergent public health challenge in African cities. High ambient temperatures can cause considerable morbidity and mortality in urban areas. Temperature gradients and geographic, socio-environmental and demographic factors determine the magnitude and pattern of health impacts. Understanding this complexity is key to developing effective responses that fit cities' spatial and demographic heterogeneity. This project, which will take place in Abidjan, Ivory Coast, and Johannesburg, South Africa, aims to be the most extensive investigation of the risks posed by heat exposure and urbanity in Africa to date. This will be in aid of developing an early warning system aimed at mitigating the impacts of heat for highly vulnerable groups[2].

We will develop and implement data science methods, ranging from natural language processing to predictive geospatial analysis, to integrate and interrogate diverse data streams alongside conventional health data. Risk is a function of exposure to a hazard, inherent vulnerability and the anticipated consequence of exposure. Thus, apart from individual demographic, economic & behavioural factors contributing to vulnerability to adverse health outcomes, we will focus on capturing vulnerability characterised by the urban form of the two African cities. Image processing, such as convolutional neural networks, will be applied to available satellite imagery to analyse urban form changes, including building types, building, street & green area densities. Census and other geospatial survey socioeconomic data will also be critical in assessing vulnerability to heat in both African cities. An important component in determining risks is hazard estimation. For this, recently developed high-resolution climate re-analysis & forecast data, meteorological station observations, and satellite imagery for inferring land surface characteristics (including land surface temperature mapping & identification of urban heat islands) will be used for historical & future heat hazard estimation for each city. To develop an early warning system that is useful for mitigating health risks posed by exposure to heat among high-risk groups, a predictive model that learns the associations between heat hazard exposure and health outcomes will be required. Health outcomes data will come from cohort and clinical trial studies conducted in Johannesburg and Abidjan.

Various approaches to disseminating information from the early warning system will be considered, including the potential development of a web-based application. The Project aligns closely with the DS-I Africa objectives, especially its solution focus, possibilities for expansion to other Research Hubs, and potential progressive expansion to cities across Africa.

# 

# Data categories

Multiple categories of data will be used across the project, broadly divided into three categories:

* 1. Health-related data These data will include data collected from previous clinical cohort and trial studies (e.g. HIV treatment trials, COVID studies), some involving pregnant women, postpartum women and their infants up to 1 years of age)
  2. ***Health-related data for RP1*** will be identified through a systematic review, mapping large (n>1000) longitudinal studies on pregnant women in sSA. These data will be acquired through direct engagement with the study PI or other appropriate study custodians or through existing data-sharing platforms that store or curate research data, such as WWARN and BMGFKi. When the data indexed on data-sharing platforms only includes the metadata or missing important variables, then the study PI would need to be contacted for primary data, in which case, an agreement with the original project/source would then be initiated. (See RP1 Protocol Document for full details)
  3. ***Health-relateddata for RP2*** will focus on datasets generated by clinical trials or cohorts undertaken in the case study cities (initially Johannesburg and Abidjan). Datasets will be identified based on the trial's scale and the availability of geospatial variables (e.g., clinic locations or other geospatial information) to allow spatial mapping of health outcomes and the intersection with socio-economic spatial mapping and climate variable spatial mapping.[[1]](#footnote-1)

## Original study Data:

This category includes raw, unprocessed data collected directly through various cohort studies and clinical trials. This data is owned by the data providers who conducted or commissioned the studies. The original study data is retained for five years following the termination of the project or the Data Transfer Agreement (DTA). Access to this data is restricted to a small team of data managers within the HE²AT Center’s DMAC. . Regarding identifiability, the source data is protected pseudonymous data, meaning it has been de-identified in compliance with the Protection of Personal Information Act (POPIA). However, it still contains indirect identifiers such as dates and addresses that, while not directly revealing personal identities, could be used for identification if cross-referenced with other information.

## Consortium Shared Data:

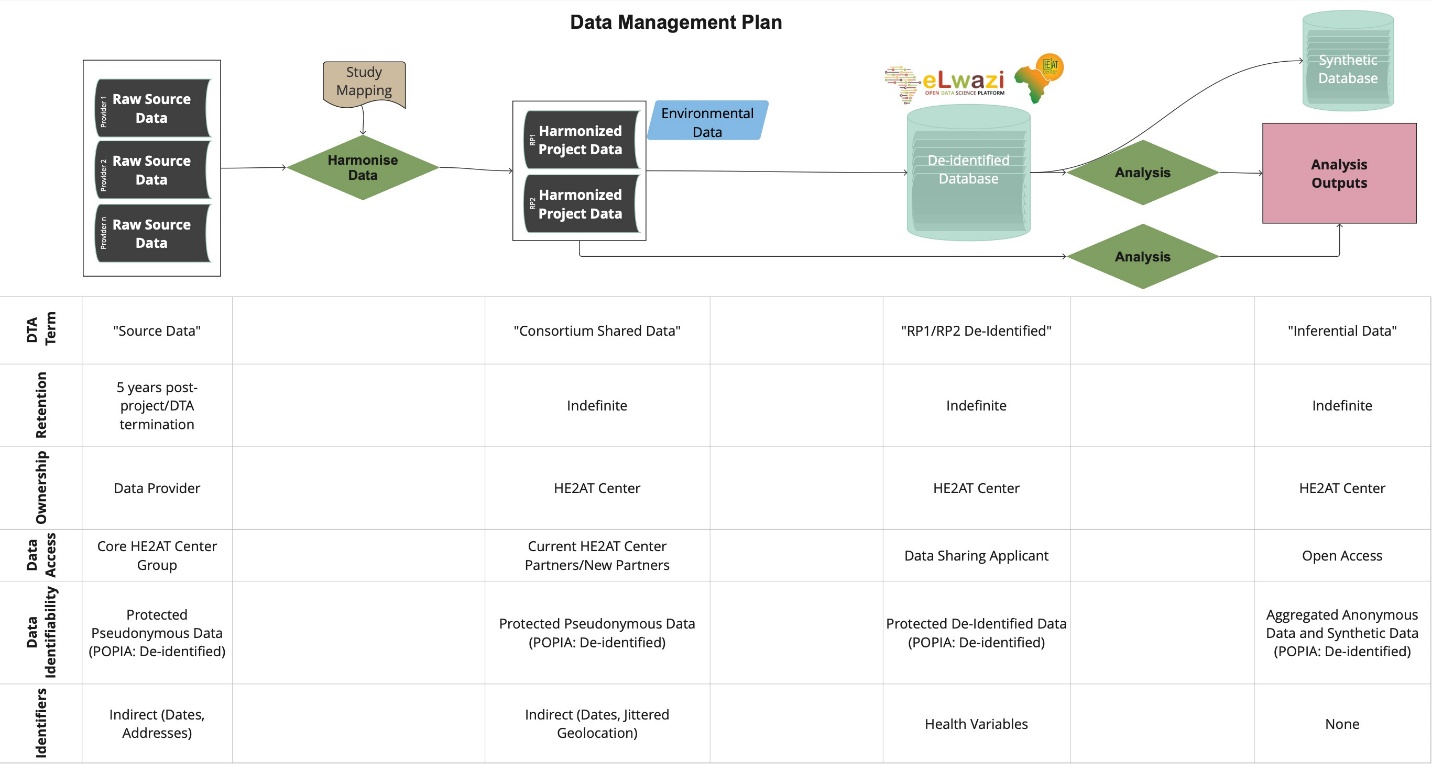
Once the source data undergoes initial harmonisation, it becomes consortium-shared data. Unless data providers state otherwise, the HE²AT Center takes control of decisions around data usage and access at this stage due to the significant alterations, harmonisation and integration work performed on the data. However, original data providers can negotiate ownership if desired. This data is retained indefinitely or for five years post-project or DTA termination, depending on the agreements in place. Access to consortium-shared data is granted to current HEAT Center partners and can be extended to new partners. This data also falls under the category of protected pseudonymous data as per POPIA guidelines, with indirect identifiers such as dates and jittered geolocation data to enhance privacy protection. NOTE IF COUNTY/PROVINCE/TOWN are included and age in decades

## RP1/RP2 De-Identified Data:

Data that has been further processed and de-identified for specific analyses related to RP1 and RP2 falls into this category. The HE²AT Center retains ownership of this de-identified data, which is stored indefinitely. Access is granted to data-sharing applicants who meet specific conditions and requirements set by the Data Access Committee. This data type is categorised as de-identified data under POPIA, meaning it has been thoroughly de-identified to prevent any reasonable possibility of re-identification. The identifiers in this dataset are limited to health variables without any direct personal identifiers. NOTE IF COUNTY/PROVINCE/TOWN are included and age in decades

## Inferential Data:

The final category is inferential data, which are aggregated and synthetic data derived from the analysis of the preceding data categories. This data is owned by the HE²AT Center and is retained indefinitely. Inferential data is made available for open access to support broader research initiatives. As aggregated anonymous data and synthetic data, it is classified as de-identified under POPIA, ensuring that individual privacy is fully protected. This category contains no direct or indirect identifiers, making it impossible to trace back to any individuals.



By structuring health-related data into these categories, the project ensures meticulous data management while adhering to ethical and legal standards. This approach facilitates collaborative research, enables new partners to join the consortium, and ensures that data privacy and protection are maintained throughout the project's lifecycle.

* 1. Climate/weather data  
     These data include observational-based datasets (weather station observations, satellite proxy observations, and processed/gridded observations). Gridded climate data from atmospheric re-analysis and climate simulations will form historical gridded climate observations and forecasts.

Climate-related data almost always involves accessing open data repositories such as Copernicus Climate Data Store (CDS) or Earth System Grid Federation data systems. Climate-related data will be stored on CSAG/UCT data storage systems; however, CSAG/UCT will manage and update the primary data index for climate-related data available to the consortium.

Most climate datasets used are available through open data policies, with no restrictions on non-commercial research use. In most cases, citing the source is required.

* 1. Remote sensing dataThe focus is on data derived from satellite sensors, mainly optical imagery (e.g., satellite images of urban centres), as well as indicators of physical measures (e.g., land surface temperature, soil moisture estimates, vegetation condition, land use and cover, etc.), to estimate environmental quantities and land use/building density, etc. Remote sensing data will not be used to identify individuals in any way and does not constitute sensitive or personally identifiable data.

Remote sensing-related data will, in most cases, involve accessing open data repositories such as Copernicus Climate Data Store (CDS), Sentinel data systems, etc. Remote sensing-related data will be stored on IBM’s data storage and CSAG/UCT data storage systems; however, CSAG/UCT will manage and update the data index for remote sensing-related data

Most remote datasets are available through open data policies, and non-commercial research use is not restricted. In most cases, citing the source is required.

* 1. Areal/Geospatial socio-economic dataThese data represent measures of socio-economic and related conditions, such as household economic status, access to services such as water and sanitation, and dwelling type. Typical sources include national census data, and more focused household and demographic survey data.

Socio-economic data will be sourced from both open data repositories and restricted-access repositories (e.g., South African census data and GCRO Quality of Life Surveys). Primary copies will be indexed and stored on CSAG/UCT data storage, but versions may already exist or can be uploaded on IBM’s system to enable analysis through PAIRS.

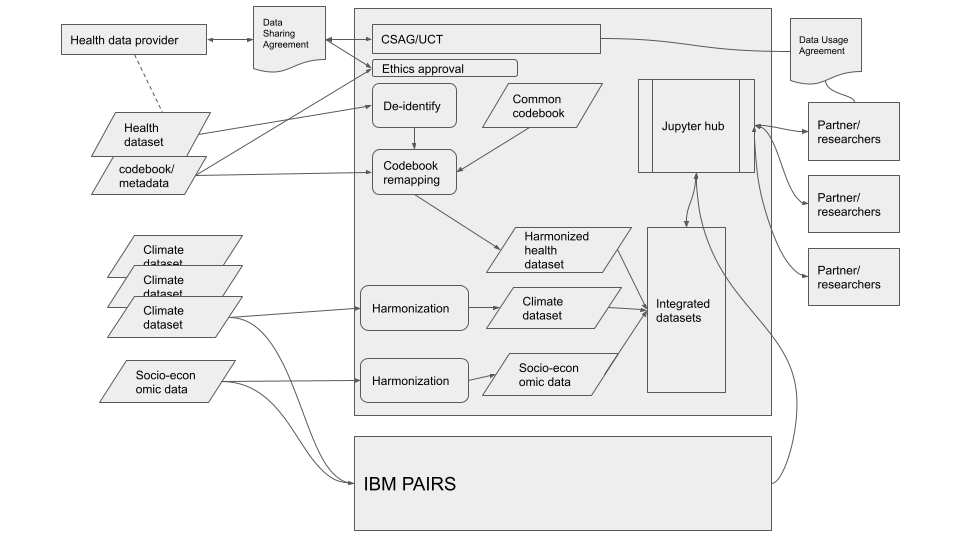
South African census data is already available through the UCT DataFirst data repository, and GCRO QoS data is available through the GCRO open data platform, which directs queries with GCRO.

South African census data is aggregated into small areas and does not constitute personally identifiable data. Likewise, GCRO Quality of Life survey data is aggregated to small areas and does not constitute personally identifiable sensitive data.

# 

# Data management workflow

Figure 1 below maps out the data processing workflow. Each element within this workflow is described below in detail.



*Figure 1: Data flow diagram indicating data providers, data sharing agreements and ethics approval, data harmonisation steps (aligning with standard formats, meta-data, etc.), integration, analysis through Jupyter Hub by partners/researchers (Diagram needs to be updated – for example IBM PAIRS system - changed*

## Setting Up the Data Transfer Agreement (DTA)

A Data Transfer Agreement (DTA) between each health data provider and UCT or WITS Planetary Health Research is required to enable health data transfer, processing, analysis, and potential publishing of research findings. The DTA is signed by the UCT or WHC legal department on behalf of the HE²AT Center and by the legal office of the data provider. In rare instances, the PI of a study may sign the DTA rather than the legal office.

## Ethics Approval for New Databases

As new databases become available and DTAs are established, the following steps will be taken to notify the Wits Health Research Ethics Committee (Medical). These processes aim to ensure that all studies adhere to ethical guidelines and processes, and maintain the integrity and compliance of the HE²AT Centre's research activities. The HEAT Center will provide additional information on the studies if requested by the ethics committee.

**Ethics Notifications for RP1 Studies**

The project notifies the ethics committee in writing of new studies that contribute data on a 6-monthly basis.

**Ethics Notifications for RP2 Studies**

1. **Data Transfer Agreements (DTAs):**
   * Prior to notification, the data-sharing agreements must be signed and fully executed with each data provider.
   * Principal Investigators must provide written permission for the use of their data.
   * Each DTA will be linked to the relevant protocol and submitted as part of the ethics notification process.
2. **Ethics Approvals:**
   * The ethics committee will review the period since the study closure to identify any potential concerns.
   * Each study will be individually assessed to ensure that ethical standards have not changed since the original approval.
   * Original ethics approvals remain valid as no new data is being collected as only secondary data analyses is being done on original data.
3. **Notification Form:**
   * A simplified, two-page notification form is provided in Appendix C, and will be completed. It includes the study background, motivation, design, and specific data fields to be collected.
   * The notification form must be submitted along with fully signed DSAs.
   * Wits HREC (Medical) will review the notifications and identify any concerns for discussion.

**Notification process:**

* + The HEAT Centre must provide the ethics committee with: a completed Notification Form for access to a new database, including background and objectives, methods (if different from the primary protocol), and a justification; and a fully executed DTA.
  + Data use may proceed once the ethics committee acknowledges the notification.
  + The study's primary objectives must be published before the HE2AT Centre publications.

## 5.3. Data encryption and transfer

Once the Data Transfer Agreement (DTA) has been agreed upon and signed and the ethics committee has raised no ethical concerns, the provider will transfer the data to the UCT data platform.

Encryption of Personally Identifiable Data:

* If the data contains personally identifiable information, it will be encrypted by the data provider before transfer.
* The encrypted data will be transferred using a secure service that employs encrypted data transport protocols, such as TLS (Transport Layer Security).
* An example of such a service is FileSender, which ensures the secure transfer of encrypted data.

This process ensures the protection of personally identifiable data during transfer, and maintains confidentiality and compliance with data security standards.

Peter to flag with Roger for updating on current process

## 5.4. Data storage and encryption

Once transferred to UCT, if the dataset constitutes personally identifiable data, or if stipulated by the DTA, the dataset will be encrypted for storage using 256-bit AES (Advanced Encryption Standard), a standard established by the US NIST (National Institute of Standards and Technology) with encryption keys only available to the minimum number of people (the Core HEAT Center Data Group) required to implement any anonymisation or data minimisation process (see team responsibilities below) (Peter – to loop in Roger for revision)

Metadata (dataset descriptions, descriptions of the original protocols, codebooks, etc.) will be decrypted and stored separately to allow for meta-data indexing and software code development.

## 5.5. Data indexing

Data currently available on the CSAG/UCT data storage system and data available on IBM’s system will be indexed using an appropriate meta-data standard, and this index will be made available by the CSAG/UCT data platform. (CKAN - Change to eLwazi in terms of sharing outside of HE²AT Data

CSAG currently implements a **Data Reference Syntax (DRS)**, a structured mapping from a controlled vocabulary of meta-data elements to a directory and file naming syntax. This is standard practice within climate data management, and we will continue to implement this approach for the climate and remote sensing datasets.

The current DRS is documented on the CSAG GitLab Wiki.

Integration with the DSI-Africa Open Data Science Platform meta-data index will ensure that meta-data propagates to the ODSP system and datasets become discoverable through ODSP meta-data queries.

Health data indexing in terms of codebook that applies relevant ontologies

## 5.6. De-identification

De-identification may be implemented (see the section on de-identification below) to minimise the potential for individuals to be identified by geospatial information or other personal information. Whether de-identification is implemented will be guided by the principle of minimalism under POPIA (not bound to POPIA – reference another more stringent standard for de-identification – eg. WHO standard), which states that *Personal information may only be processed if, given the purpose for which it is processed, it is adequate, relevant and not excessive* (POPIA Chapter 3: Section 10 Minimality). For example, for some research analyses, street address level personal information is not required and will be replaced by larger area references such as South African census areas.

* Data shared for analysis only available after “geographic jittering” not actual addresses
* Move the bottom "de-identification" to here!

## Codebook remapping and harmonisation (Peter – Renaming this section)

## Health Data Harmonisation

Data harmonisation is critical for integrating diverse datasets into a unified format, allowing comprehensive analyses drawing on information across the disparate data sources. A detailed description of the workflow from source data to deidentified data is as follows.

1. Pre-Processing:

This initial step is done by the ‘core’ HE²AT team (definition – those with access to ‘incoming’ (peter, pierre, lisa, chris, rodger, nic, craig)) and consists of:

- reformatting to common data format

- extracting variable names and descriptions

- creating synthetic data (ie 10 random samples of each variable)

- identify relevant study documentation

This data is periodically backed up to a secure and encrypted disk seperate to the wider HPC system. This data is not encrypted but access permissions are restricted to the ‘core’ group.

1. Variable Mapping

This step is completed by the full consortium making use of synthetic data, variable metadata and study documentation created in step 1. This data is shared with OpenAI to take advantage of LLM description completions, variable mapping suggestions and data transformation cleaning/ suggestions. At this step harmonisation team members complete a form to map source variables to a standard set of ontologies and recode variables to a standard data model.

1. Mapping Validation

This step is completed by the ‘core’ HE²AT team making use of the variable mappings created in step 2 as well as the reformatted source data in step 1. Here a second validation of the mapping and transformations are completed first by a health expert and second by the core harmonising team. The results from this and step 2, referred to as mappings and transformation instructions are version controlled using the CSAG Gitlab.

1. Database Population

This step makes use of the mappings and transformation instructions created in step 1 and 2 to first transform the source data and then populate a final database. This step is managed by the ‘core’ team. The final database is available to the full consortium. Note a second anonymisation of patient identifiers will happen at this stage and locations below the ward level (ie street addresses) will not be included.

1. Final Data Cleaning and Analysis Dataset Creation

A second data cleaning step will be performed on the database to ensure values do not fall outside of expected ranges, provisions are made to identify twins and multiple pregnancies, and duplication errors are identified, and a final sanity check will be performed to ensure data integrity. Linking geospatial data will be an additional step performed at the analysis stage.

## Data integration and analysis (Revise – integrate w/above section; focus on the development of integrated relational database)

Integrated datasets (with associated documentation provided through the GitLab platform) will be available for analysis through the Jupyter Hub platform to partner researchers. In some cases, depending on the conditions of the original Data Transfer Agreements (DTAs), these partner researchers will have to agree to specific data usage agreements.

The complete health data processing workflow is mapped in Figure 2 below. The key steps are:

1. **Basic Error Screening and Correction:** Identifying and correcting formatting errors, invalid codes, and values.
2. **Screening for Personal Identifying Information:** Checking for personal identifiers such as date of birth, which can be a high risk for re-identification. For example, date of birth can be dropped after calculating and retaining the patient's age.
3. **Recoding to Common Codebooks:** Extracting variables of interest for each RP, translating variables into the target units, including conversions of real values into categories and translating between categorisation schemes. This step also involves recording geographic information. For example:
   * **Point Location Data:** Using jittering to perturb spatial location data systematically (random distance and direction) or substituting point locations with grid-cell centroids corresponding to the grid framework used for climate and environmental data.
4. **Further Quality Control and Data Restructuring:** Ensuring the accuracy and consistency of the data before integration.
5. **Integration of Climate Variables:** Incorporating climate data to support specific analyses.

By standardizing variables and conforming to established meta-data and storage standards, this process ensures that the combined dataset is robust, reliable, and suitable for advanced analytical techniques. This harmonized dataset will serve as a critical resource for ongoing and future research within the HE²AT Center and beyond.

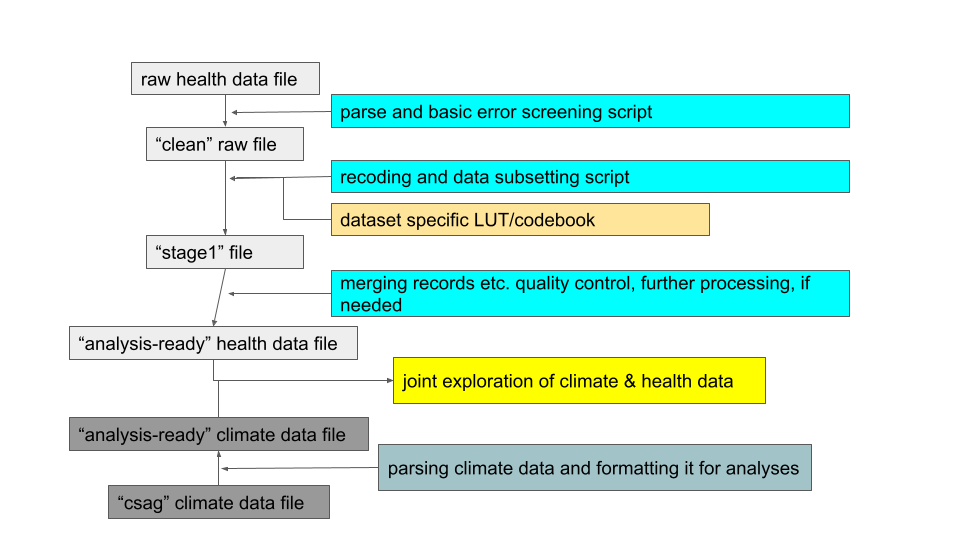


Figure 2: Detailed data processing for health data ingestion

## Data analysis platforms

Data analysis will be enabled through the CSAG/UCT Jupyter Hub as well as IBM:

*Jupyter Hub* is a collaborative web-based Python coding environment that enables analysts to develop analysis code in Python using a web-based coding platform. The analysis code is executed on the CSAG HPC platform and datasets stored on the CSAG storage servers are available.

*IBM* provides multiple interfaces for rapid and more complex analysis, which will be made available to HE²AT Center analysts on request (see Roles and Responsibilities for access request contacts) Sibu to update. In addition, approaches to integrate data from IBM into the CSAG/UCT Jupyter Hub will be developed.

# Data Management, Documentation and Curation

Data will be managed by the DMAC team **[DMAC@witsphr.co.za](mailto:DMAC@witsphr.co.za:%20Sender)**

and ***primary data management*** will occur on the CSAG/UCT data platform. (Just refer to DMAC team as described elsewhere in doc – check with Lisa)

The data management practices that maintain the IBM systems will continue. Critically, for synergy with the CSAG/UCT platform, meta-data indexing will be shared with CSAG/UCT for data related to the NIH HE²AT Center and DS-I Africa. IBM will not host health data, so data privacy and access restrictions will be limited to data accessed through the CSAG/UCT platform.

Data upload to the CSAG/UCT platform from partner institutions should follow procedure including proper data cataloging, documentation, associated meta-data files, user access groups and citations. Potential to consider alternate, secure data sharing platforms such as cloud object storage for larger climate datasets. In the case of supplementary projects with external partner institutes (no DTA with HE²AT Center, e.g. CHEAQI partner MSU), user access groups need to be defined before data is transferred to CSAG/UCT platform following the above process..

Individual partner institutions will have data and computing platforms, ranging from central institutional platforms to personal computing devices. Data management within these partner institutions does not fall under this data management plan, and we refer to it as ***secondary data management***. However, partners will be encouraged to regularly consider what locally managed data should be integrated into the project-wide data management platform and made available to the broader DSI-Africa programme and beyond. Health data must be kept in a secure and managed platform; therefore, where researchers from the various partner institutions wish to run analyses on the health data in the CSAG/UCT platform, they will use the DMAC JupyterHub environment. (Missing – procedure for data sharing / transfer amongst HE²AT partners – requirements, formats, elements of metadata file)

Primary data management will involve:

* Homogenization to agreed data archive standards (see above for each class of data)
* Source data is backed up and remapping/recoding instructions are version controlled Harmonizing of health-related data
* Meta-data indexing within DS-I Africa ODSP
* Documentation of data on CSAG/UCT GitLab wiki
* Sharing of data management code through CSAG/UCT GitLab code repositories

# POPIA compliance and protection of personal information

The use of health datasets requires careful consideration of data security and confidentiality. This is strongly guided by the relevant legislation context for each dataset, including the specific country legislation around using personal/sensitive data and the cross-border transfer of such datasets. The development and negotiation of these data-sharing agreements lie in the intersection of DMAC and research projects, as the research projects are the primary interface with the data sources.

The Protection of Personal Information Act of South Africa (2013) institutes limitations on processing personal information but also provides the legal basis for using personal information for scientific research. POPIA functions alongside other legislation and regulatory structures governing research in South Africa, such as the Constitution of the Republic of South Africa, the National Health Act No 61 of 2003 and regulations as prescribed by the Minister of Health, such as the South African Department of Health guidelines on Ethics in Health Research Principles, Processes and Structures 2015 (“DOH Guidelines”). The law that takes precedence will provide the most comprehensive protection for the rights of individuals in South Africa.

Section 6 of the Act indicates that the Act does not apply in cases where personal information has been de-identified to the extent that it cannot be re-identified again. Where this is the case, the information can be used without restriction or conditions of the Act. It is envisaged that a number of health databases acquired for HE²AT will have been de-identified to the extent that re-identification is virtually impossible. However, where this is not the case, the following sections of POPIA provide a basis for processing health data.

While Section 15(1) requires that personal information processing must be compatible with the purpose of collection, **Section 15(3)(e) allows for processing for the purposes of historical, statistical and research purposes** regardless of the original purpose of collection.

As HE²AT will be using health datasets collected before the project, we will never be processing personal information compatible with the original purpose of collection, unless the data subjects agreed to specific consent for further processing. Section 15(3)(e), therefore, provides the basis for processing information where this specific consent was not obtained.

Furthermore, while Section 18(1) lays out requirements for informing the data subject that their personal information is being processed for a particular purpose, **Section 18(4)(f) provides for an exemption from these requirements where the information will be used for historical, statistical or research purposes.**

Furthermore, **Section 14(2) allows for the retention of personal information for research purposes** as long as safeguards are in place to prevent the use of the personal information for other purposes (see data encryption and storage isolation below)

Section 16 requires that the responsible party takes reasonable measures to ensure data quality is maintained. **We will implement data quality control as per section 8 above.**

Section 17 requires that clear documentation regarding all processing is maintained. **Our code management and meta-data management plans provide for this.**

Section 19 requires that security measures are implemented to prevent the unlawful access to or processing of personal information. **We will implement extensive data security measures (see data encryption and storage isolation below)**

Section 20 describes the requirements for individuals operating on or processing personal information. We have identified the individuals responsible for this processing and will continually update this list. **Access to personal information is restricted through passwords and other security measures so that only individuals authorised by the responsible party have access.**

Section 21 requires a written contract between the responsible party and the operators implementing processing. This contract will be signed by all individuals processing personal information and is presented in Annex 1 and includes agreement that the operator will inform the responsible party where there is any basis for believing that personal information has been accessed by an unauthorized person.

# De-identification

POPIA Section 10 prescribes the principle of Minimality, which means that only information relevant to the purpose of processing should be processed. Where personal information is acquired that is required to fulfil the research purposes described by the relevant research project protocols, de-identification will be implemented according to the following protocol, which is guided by US Department of Human and Health Services (HSS) guidelines and informed by the findings in Zandbergen’s 2014 review on geographic masking strategies[[2]](#footnote-2).

## 10.1. Geographic Aggregation:

* + Street addresses may be aggregated into larger geographical regions to prevent the derivation of individual residential locations. For instance, in RP2, where the highest spatial granularity is needed to map urban heat-health outcomes, data will be aggregated at the level of census small areas or wards with spatial scales of 2 to 5 km. This approach ensures that many records map to the same region, mitigating the risk of re-identification. The aggregation process will also consider the number of records that map to the same geographical area to ensure privacy.

## 10.2. Location Jittering:

* + Latitude/longitude coordinates may be "jittered" by adding random values to each coordinate to obscure the exact location while retaining sufficient geographical information to support analysis. As detailed by Zandbergen (2014), jittering can involve various methods:
    - **Random Direction and Fixed Radius**: Masked points are placed randomly on a circle around the original location, ensuring a fixed displacement.
    - **Random Perturbation within a Circle**: Masked locations are placed within a circular area around the original location, with the displacement distance following a uniform or normal distribution.
    - **Gaussian Displacement**: The direction of displacement is random, but the distance follows a Gaussian distribution, with dispersion adjusted based on local population density.
    - **Donut Masking**: This technique sets a minimum and maximum displacement level, ensuring masked locations are neither too close nor too far from the original points.
    - **Bimodal Gaussian Displacement**: A variation of Gaussian masking, employing a bimodal distribution to achieve similar effects as doughnut masking but with a less uniform probability of placement.

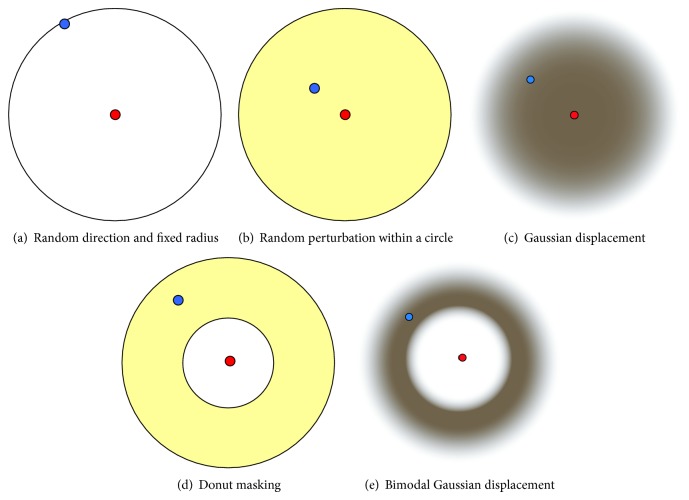


Figure 1: Geographic Masking Techniques: Different geographic masking techniques as described by Zandbergen (2014). (a) Random direction and fixed radius; (b) Random perturbation within a circle; (c) Gaussian displacement; (d) Donut masking; (e) Bimodal Gaussian displacement.

* + For RP1, jittering will be used to shift latitude/longitude locations by tens of kilometres, adequate to prevent locating individual residential locations. This approach will involve UCT, IBM, and NIH expert determination to ensure the balance between data utility and privacy.

1. **Expert Review and Validation**:
   * Geo-location masking/jittering and aggregation techniques will be reviewed and validated through expert determination involving UCT, IBM, and NIH experts. This process will involve assessing the risk of re-identification and ensuring that the applied techniques sufficiently protect participant confidentiality while maintaining the integrity of spatial analyses.
2. **Assessing Re-identification Risk**:
   * The risk of re-identification will be quantified using spatial k-anonymity metrics, as described by Zandbergen (2014). This involves ensuring that each masked location is indistinguishable from at least k-1 other locations within a specified distance. The displacement required for adequate masking will be inversely proportional to the local population density to maintain high spatial k-anonymity.

By incorporating these enhanced de-identification techniques, we aim to ensure compliance with POPIA, protect participant privacy, and maintain the utility of the data for research purposes.[[3]](#footnote-3)

# Data encryption

Primary data (before processing) will be stored in encrypted form using 256bit AES (Advanced Encryption Standard) a standard established by the US NIST (National Institute of Standards and Technology) with encryption keys only available to the minimum number of people required to implement the de-identification processing.  
  
The use of cryptographic modules validated to NIST [FIPS 140-2](https://en.wikipedia.org/wiki/FIPS_140-2) is required by the United States Government for encryption of all data that has a classification of Sensitive but Unclassified (SBU) or above.

For datasets that include personal identifiers (ie. personally identifiable data), datasets will be encrypted by the original data holder and transferred to UCT through a secure data transfer that uses TLS encryption.

# Storage isolation

Encrypted data will be stored in an isolated virtual storage server that is only accessible to the private data team.

# Network firewall and Virtual Private Network

CSAG compute infrastructure is part of the UCT intranet and is covered by the broader UCT security policies and services. This ensures that network access to CSAG servers and computer services is carefully controlled and limited to authorised users only.

*UCT broader security policy*

<http://www.icts.uct.ac.za/sites/default/files/image_tool/images/286/UCT_Information_Security_Policy_PC03_2020.pdf>

UCT ICTS implements a Cisco firewall and Cisco Virtual Private Network service to ensure that intranet access is limited to authorized UCT users only.

The UCT firewall policy describes the UCT firewall implementation, which ensures that only authorised access to the UCT intranet occurs.  
<http://www.icts.uct.ac.za/uct-perimeter-firewall-policy>

These network-level security measures ensure that access to the UCT intranet is controlled by authorised UCT users. Furthermore, CSAG server access is controlled by CSAG authentication and authorization as described below.

# Local authentication and authorisation

While CSAG compute servers sit within UCT's intranet, the CSAG/UCT platform (storage and compute infrastructure) implements an independent authentication and authorization service (Linux filesystem and LDAP authentication). Access to restricted datasets will, therefore, be ensured through both UCT authentication protocols and internal CSAG DMP authentication and authorizations.

# Data Sharing and Open Access

According to the Research Data Management policy, ‘publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner’. Data is, therefore, open by default and closed by exception (e.g. privately funded research or research with commercialisation possibilities).

**Restrictions to data sharing**

According to Section 4.6 of the [UCT Research Data Management policy](http://www.digitalservices.lib.uct.ac.za/dls/rdm-policy),

"[n]ecessary constraints on the availability of data include the protection of personal data; the protection of intellectual property; the protection of commercial interests of project partners; and security concerns."

**Discoverability**

The HE²AT Center and DMAC will implement FAIR principles:

* Data will be **findable** through publicly accessible and searchable meta-data indexes (need to decide if the DS-I Africa ODSP is going to be the primary mechanism for meta-data search; UCT also hosts the ZivaHub repository)
* Data will be **accessible** either openly through a public-facing component of the DMP data repository or through a data access request to DMAC (where a Data Sharing Agreement is required)
* **Interoperability** will be enabled through the strict adherence to established data and meta-data standards (see above)
* **Reuse** will be supported through rigorous data documentation, including limitations and guidance for reuse.

# Procedure for making data available to qualified individuals

The Steering Committee will develop a Data Access Request Form for researchers outside of the HE²AT Center team to request access to the RP1/RP2 de-identified data. The completed form is then assessed by the Data Access Committee considering their request. The form will include a proposal outline of the intended research and the procedures to maintain data confidentiality and security. The data sharing will be conditional on proof of ethics approval, a.

The Steering Committee will review and approve or reject all research community requests, including those of scientists or medical professionals working at academic, or non‐profit institutions. Approval can be conditional on an applicant submitting their ethics approval at some stage. Data will only be shared once applicant has secured formal ethical approval for their institutional ethics committee. Data will only be shared if the applicant states that it is in full compliance with the law of its country.

We will ensure that all requests conform with NIH policies and procedures, including compliance with informed consent procedures if relevant and any limitations stipulated by the institutions/investigators who contributed data to either RPs.

Researchers who request to share the resources of the HE²AT Center will need to agree not to seek to identify individuals within the dataset, not distribute the data to any other entity, keep the data secure, and acknowledge the HE²AT Center and DS-I Africa as appropriate in publications and presentations (the exact acknowledgement text to be agreed among the DS-I Africa Program). Further, researchers who share the Center’s resources will be strongly encouraged to collaborate with and train African investigators as part of their work with the resources shared.

There may be exceptions to the resource-sharing plans outlined above. Firstly, there may be considerations around intellectual property protections for the research products that the consortium aims to commercialize. Decisions about resource sharing in these circumstances will follow the NIH policies in this regard, including those on resource sharing, disseminating unique research resources, and program income[1, 2]. The NIH will receive a copy of documents or samples of these products developed under the grant award (e.g. the Digital App). A large portion of the data we will use in the Project activities will be drawn from the IBM platform, which contains several datasets that require a license for access. We will thus not be able to share those datasets without permission from the licensee.

Applicants will be contractually bound not to attempt to reidentify individuals. The Data Access Committee may remove further indirect identifiers on a case-by-case basis, based on a risk assessment of the applicant.

The rights and privacy of individuals who participate in research must be protected at all times. Thus, data intended for broader use should be free of direct identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects. All data shared beyond the HE²AT Center will be de-identified following the abovementioned procedures. Requests for original data containing personally identifiable or sensitive information will be referred back to the original study.

The HE²AT Center team will especially endeavor to make the unique research resources we develop readily available to other researchers working on climate and health. There is a pressing need within the field to expedite data translation into knowledge, particularly interventions that protect people against extreme heat and other manifestations of climate change. We thus feel obligated to expedite our responses to such data requests. Fellow researchers wishing to access the data will thus not have to wait for our research findings to be accepted for publication or the ‘final research data” before data sharing, provided the research questions they are exploring do not directly overlap with the specific question we are addressing.

The IPD in RP1 involves ‘data sharing’ on a large scale, where we rely on data providers' willingness and ability to share. In that spirit, as the owner of the IPD RP1/RP2 Deidentified database, it will be our aim to share the data with other groups, provided the data providers have agreed to share data in the DTA, and the procedures of HE²AT Center data sharing processes are followed.

[CHECK IN DTA THAT IF THEY OPT OUT OF ON SHARING THAT THE OPT OUT ONLY AFFECTS ORIGINAL SOURCE DATA, OR DEIDENTIFIED DATA..]... we will adhere to whatever level of sharing that the provider has agreed to...

We will develop formal collaboration agreements around data sharing with the data providers who contribute data to the IPD, which sets out the terms and conditions for data reuse. This is important, as some of these investigators may not wish their data to be shared with third parties. While a willingness to share data forms the basis for an IPD and most often reflects a desire of investigators to collaborate, this may not necessarily translate into a willingness to share data beyond the Center then. Where data providers have not agreed to sharing deidentified data to external parties in the signed DTA, their data will not be shared. The synthetic datasets and the aggregated datasets that we create using their data can be shared.

Sharing the datasets gathered into the IPD database will thus require a signed agreement from the original data owners and may require an application to the local IRB where the original study was done.

People who use the IPD database must adhere to the authorship guidelines stipulated in the collaboration agreement signed between the HE²AT Center and the research groups who contributed their data.

It is worth noting that much of the data and resources generated by the Hub will be helpful in a range of other disciplines. The research resources generated by the RP2 team, such as vulnerability-heat-health data visualisation, will also potentially have a wide range of applications among people working on urban geography or planning. Additionally, the datasets we will generate in the IPD in RP1 have tremendous potential to answer a wide range of questions outside of environmental health. We will facilitate interested parties' access to these resources. We also share data and resources generated by the HE²AT Center with other Hubs or components in the DS-I Africa program, wherever possible. In particular, the DS-I Africa Open Data Science Platform will be used to make relevant datasets from the HE²AT Center available across the program and beyond.

In terms of the DMAC activities, data and data resources (primarily software/code) will be shared through standard protocols such as File Transfer Protocol (FTP) servers and the UCT instance of Gitlab (software version control system). UCT also hosts an open data portal, which will be used to make more final datasets, such as the urban heat vulnerability maps developed in RP2, available following FAIR principles[3]. Data that is made publicly or otherwise available through data use and sharing agreements will be accompanied by metadata such as data dictionaries and data descriptors such as principal investigator, funding sources, data collector, project description, sample and sampling procedures, temporal and geographic coverage of the data collection, variables, technical information on files (file formats, linking, etc.), interviewer guides and coding instruments.

Analytic data sets will be provided as de-identified data files that can be read by common statistical package software, such as SAS or Stata. De-identified data sets will have names and other personal health identifiers removed (see above). Data and other resources will be transferred to others under the terms of a data-sharing agreement to ensure that the data will be used for the proposed purpose and that no attempts will be made to identify participants. We will maintain records of all researchers who are given access to the research resources. Results of the HE²AT Center activities will be shared with the research community and the public through conference presentations, publication in peer-reviewed journals and media interactions, as described in the Training and Engagement Core.

# Governance of data sharing

**Draw on the Data Sharing and Access Guideline for DS-I Africa; and other DACs**

Protection measures

DPIA

Make public the data we have

# Data Access Committee

The Data Access Committees (DACs) plays a crucial role in the governance of data access within the DSI Africa Consortium. It evaluates and decides whether and on what conditions to make data available to bone fide researchers that apply for data. The DAC ensures that data access requests are evaluated fairly and in alignment with the ethical guidelines, privacy standards, and research objectives.

Following completion of the HE²AT Center funding, data oversight will need to be maintained either through continuation of the existing DAC provided funding is available; or through the DS-I Africa.

Where it is the case that the HE²AT Center does not own the data instances in the Consortium-Shared data, these data will be destroyed.

**Membership of the DAC:** UCT ethics representative; ELSI team representative; eLwazi representatives; Study Principal Investigators; independent scientists from outside the HE²AT Center.

**Terms of reference**:

We follow a protocol in reviewing data access requests, outlined in the steps below:

1.1.2 Step 1: Confirmation of Bona Fide Researcher Status

External applicants, i.e., those not affiliated with the HE²AT Center Consortium, undergo a thorough verification to confirm their status as bona fide researchers. A DAC reviews the provided credentials, affiliations and the research track record to ensure that applicants are engaged in legitimate scientific research and adhere to ethical research practices. This may involve requesting additional documentation or references to substantiate the applicant's bona fide status.

1.1.3 Step 3: Assessment of Research Purpose

A DAC applies three criteria for access:

First, that the data request must be aligned with the Data Use Permission (see clause 11.2.5 above).

Second, that potential risks to the privacy of research participants associated with the proposed use of the data, where relevant, have been identified and addressed in a satisfactory manner.

Third, that the proposed research does not overlap with or pre-empt the results of the ongoing Consortium projects. Should the latter be the case, the DAC should impose a temporary embargo on the Data Sharing, instead of declining access. An embargo period allows Consortium members to complete their research and publish results before the data is made available to external parties. The length of the embargo should be determined based on the specific circumstances and projected timelines of the involved Consortium projects.

1.1.4 Step 4: Recording and Communication of Decision Reasons

Once a decision has been made, the DAC meticulously records the reasons for its approval, conditional approval, or denial of the data access request. This documentation ensures transparency and provides valuable feedback to applicants. The decision, along with the reasons and any conditions attached (such as an embargo period), is communicated to the applicant in a clear and timely manner. For approved requests, the communication includes detailed instructions for accessing the data and any requirements or conditions that must be met by the applicant.

# Data Retention

Participant data will be retained for at least five years after the completion of the project for historical, statistical, or research purposes as provided for by POPIA. Appropriate safeguards will be established to secure the records and ensure that the data is not used for any purpose other than the purpose for which it was originally collected. In any event, no records will be retained for longer than as per applicable regulations.

# Roles and Responsibilities

The table below details the various roles and responsibilities associated with the data management plan, as well as who is currently associated with each, their institution, and contact details

|  |  |  |  |
| --- | --- | --- | --- |
| **Role and responsibilities** | **People** | **Institution** | **Contact** |
| **DMAC PIs**  Responsible for ongoing (quarterly) assessment of data management and changes to the data management plan (annual) | Christopher Jack  Sibusisiwe Makhanya | UCT  IBM | [cjack@csag.uct.ac.za](mailto:cjack@csag.uct.ac.za)  sibusisiwe.makhanya@ibm.com |
| **Health data acquisition**  Identification of relevant health datasets, coordination and development of the DTA | Matthew Chersich for RP2  Craig Parker for RP2  Stanley Luchters for RP1 | WITS PLANETARY HEALTH RESEARCH  The Centre for Sexual Health and HIV/AIDS Research Zimbabwe (CeSHHAR Zimbabwe) | [matthew.chersich@tcd.ie](mailto:matthew.chersich@tcd.ie)  [Craig.parker@witsphr.org](mailto:Craig.parker@witsphr.org)  stanley.luchters@ceshhar.co.zw |
| **Data processing and harmonisation (including de-identification)**  De-identification, quality control, remapping, harmonisation and integration of all datasets  Note: These are the only individuals with access to encryption keys for original sensitive data | [Lisa van Aardenne](mailto:lisa.vanaardenne@uct.ac.za)  [Pierre Kloppers](mailto:pierre.kloppers@uct.ac.za)  Piotr Wolski  Peter Marsh  Nicholas Brink  Craig Parker | UCT  UCT  UCT  UCT  WITS PLANETARY HEALTH RESEARCH  WITS PLANETARY HEALTH RESEARCH | [lisa@csag.uct.ac.za](mailto:lisa@csag.uct.ac.za)  [pierre@csag.uct.ac.za](mailto:pierre@csag.uct.ac.za)  [wolski@csag.uct.ac.za](mailto:wolski@csag.uct.ac.za)  Peter.marsh.uct.ac.za  nicholas.brink@witsphr.org  [Craig.parker@witsphr.org](mailto:Craig.parker@witsphr.org) |
| **Managing access to UCT data analysis platform** | Rodger Duffett | UCT | rodger@csag.uct.ac.za |
| **Managing access to IBM platform** |  | IBM |  |

# 

# Assessment and revision

The DMAC co-PIs will periodically re-assess the Data Management Plan in consultation with the HE²AT Center Steering Committee, including the RP1 and RP2 leads. This will take place at least every six *months*.

The assessment will look at three aspects of the data management plan:

1. Data acquisition processes: Are the data acquisition processes working concerning developing DTAs, transferring the data, and satisfying ethical reviews
2. Data process: Is the data processing workflow working effectively and resulting in data that is ready for analysis?
3. Data analysis support: Is DMAC providing sufficient support and services to enable the data analysis plans for RP1, RP2, and any pilot projects?

The Data Management Plan will be updated and changed based on the assessment. The DMAC co-PIs (see above) will propose revisions to the plan, which will be approved by the SC.

# References

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in medicine*, *2014*.

# Annex 1: Key data sources

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name and source of dataset** | **Description** | **UCT /**  **PAIRS** | **Key variables** | **Spatio-temporal coverage** | **Relevance** |
| **Biomedical data** | | | | | |
| Individual Participant Data platform | Collation of prospectively collected high-quality data from of pregnant women & and/or neonates  (PROSPERO:  CRD42020214637) | Data  owners; ki platform  and NICD  repository | Preterm birth, pre-eclampsia, neonatal admission | African cohorts  and trials  conduced  between 2000 and Oct 2020 | Research Project 1 |
| HIV databases | Pooled health database from multiple large HIV trials  conducted among adults in Johannesburg, South Africa | WHC  studies | Participants are followed up every 3-months for several years, with a multitude of physical measurements, laboratory tests, images and health  questionnaires |  | Research Project 2:  The study population has high rates of co  morbidities and adverse health outcomes |
| **Climate/weather data** | | | | | |
| European Centre for Medium-Range  Weather Forecasts (ECMWF) -  https://www.ecmwf.i nt/en/forecasts/data sets/set-i | Outputs from a numerical weather prediction system, run twice daily, designed to produce state-of-the-art  medium (10 days) global forecasts (contains only the latest forecast) | PAIRS | Temperature (Ground, Min, Max) at 2 m above ground; Solar irradiance; Wind speed (toward east, north) at 10 m above  ground; Daily precipitation (total, rate); Dewpoint; Pressure | Spatial: Global  coverage,  0.065536 deg.  Temporal: 3 – 6  hourly & daily res.; Jan 2014 – Oct  2019 | Determination of heat hazard; Thermal  comfort metrics;  combined climate  exposures (forecasts) |
| IBM TWC (The  Weather Company) Current and  historical weather | Data layers from The Weather Company, an IBM Business | PAIRS | Temperature (Change, Min, Max, Feels like); Solar irradiance; Wind (speed, gust & dir.), Rel. Humidity, Daily precipitation (total, rate); Dewpoint; 3-hrly Pressure Change | Spatial: Global  coverage, 4km  landmass and  coastal waterways (hourly & daily res from 2015) | Determination of heat hazard; Thermal  comfort metrics;  combined climate  exposures (historical) |
| Fifth-generation  ECMWF high-res. Reanalysis (ERA5) https://cds.climate.c opernicus.eu | A global reanalysis dataset combining observed data with the output of meteorological models. | PAIRS | Temperature (2 m above ground, Min, Max); Wind speed (toward east, north); Daily precipitation (total, rate, type); Atmospheric water/ water vapour content, Thermal radiation; Soil  temperature; Vegetation types and cover (high, low) | Spatial: Global  coverage,  0.131072 degrees PAIRS resolution (raw: 0.25 deg.)  Temporal: hourly; coverage from Jan 1980 – Jun 2019 | Determination of heat hazard; Thermal  comfort metrics;  combined climate  exposures (historical) |
| Fifth-generation  ECMWF high-res. Reanalysis ERA5-Land  https://cds.climate.copernicus.eu/cDTApp#!/dataset/reanalysis-era5-land?tab=overview | A global reanalysis dataset combining observed data with the output of meteorological models. contains hourly data from 1950 to present | UCT | Includes a range of surface and near-surface variables including: 2m temperature and dewpoint temperature, surface skin temperature, precipitation, near-surface winds, surface net thermal radiation. | Spatial: Global  coverage,  0.1 deg)  Temporal: hourly 1950 - present | Determination of heat hazard; Thermal  comfort metrics;  combined climate  exposures (historical) |
| WATCH Forcing Data methodology applied to ERA5 (WFDE5)  [https://cds.climate.copernicus.eu/cDTApp#!/dataset/derived-near-surface-meteorological-variables?tab=overview](https://cds.climate.copernicus.eu/cdsapp#!/dataset/derived-near-surface-meteorological-variables?tab=overview) | A global bias-corrected reconstruction of near-surface meteorological variables drive from the ERA5. | UCT | includes a range of surface and hear-surface variables including: near surface air temperature, specific humidity, rainfall, wind speed, air pressure and surface longwave and shortwave radiation | spatial: Global land  Temporal: Hourly 1979 - 2019 | Determination of heat hazard; Thermal  comfort metrics;  combined climate  exposures (historical) |
| Temperature and precipitation gridded data for global and regional domains derived from in-situ and satellite observations  [https://cds.climate.copernicus.eu/cDTApp#!/dataset/insitu-gridded-observations-global-and-regional](https://cds.climate.copernicus.eu/cdsapp#!/dataset/insitu-gridded-observations-global-and-regional) | Temperature and precipitation from different datasets including: GISTEMP, Berkeley Earth, CPC and CPC-CONUS, CHIRPS, IMERG, CMORPH, GPCC and CRU | UCT | precipitation, maximum, mean and minimum temperature | Spatial:  global, quasi-global, Africa depending on the dataset.  Temporal: daily or monthly depending on the dataset | Determination of heat hazard; Thermal  comfort metrics;  combined climate  exposures (historical) |
| Copernicus S2S  seasonal forecast data | Model outputs forecasting climate conditions over the three months following the forecast initialization | UCT | Temperature 2m above ground (min, max), Daily precipitation (total) | Temporal: daily | Seasonal (weeks to 3 months) time horizon forecasting of relevant weather conditions  (heat hazard) for early warning |
| CP4-A (NERC  JASMIN) | Very high resolution (4km) simulations of historical and future climate over Africa | UCT | Temperature 2m above ground (min, max), daily precipitation (total), multi-level circulation | Temporal: hourly | Dynamical downscaling to support sub-urban temp hazard mapping |
| CORDEX Africa  (ESGF) | Ensemble of dynamically downscaled simulations of African climate to 50km,  25km, and 10km resolution | UCT | Temperature 2m above ground (min, max), daily precipitation (total), multi-level circulation fields | Temporal: daily  and sub-daily (6 hourly) | Dynamical downscaling of climate to support sub-urban temperature hazard mapping |
| GHCN station data (NOAA GHCN) | Global archive of daily  weather station data | UCT | Temperature 2m above ground, daily precipitation (total) | Temporal: daily, station locations | To support statistical downscaling of  temperature hazard |
| **Remote sensing data** | | | | | |
| 30 m res Elevation (SRTM) (NASA  https://www2.jpl.nas a.gov/srtm/) | Global elevation data from the Shuttle Radar Topography Mission (SRTM). | PAIRS | Elevation | Released in 2013 | Determination of heat hazard; Urban heat  Island Effect |
| High res imagery  (ESA Sentinel 2)  European Space  Agency  https://sentinel.esa.i nt/web/sentinel/senti nel-data-access | Images from the Sentinel 2 satellite pair which view land surface regions in 13 spectral bands. | PAIRS | Urban land cover – vegetation coverage, morphological  features, possibly pollution levels (AOT). Bands 4 (red), 8 (NIR) and SCL (Scene Classification); Aerosol Optical Thickness; NDVI sh layer. | Spatial: Global  coverage;  0.000064 deg res Temporal: every 5 days or faster;  from Aug 2015 – Nov 2020. | if there is requirement to control for pollution effects or to look at  combined heat-pollution exposures |
| Aqua MODIS Land Surface Temperature (MYD21A1D - <https://doi.org/10.5067/MODIS/MYD21A1D.061> & MYD21A1N - <https://doi.org/10.5067/MODIS/MYD21A1N.061>) | Satellite derived day and night time, high resolution (1KM) land surface temperature dataset. | UCT | Land surface temperature | Spatial: Global land surface coverage; 0.00983 deg res. Temporal: daily; 2002/07/04 to present |  |
| Gauteng City-Region Observatory.  https://gcro1.wits.ac.za/gcrojsgis/ | GIS raster and shapefiles for the Gauteng City-Region area | UCT | Demographics, economics, environmental, spatial structure, spatial change and transport | Spatial: Gauteng city-region.  Temporal: various depending on the variable | Research Project 2: provides information on socio-economic circumstances and attitudes of residents within the Gauteng City-Region. |
| **Areal / geospatial socio-economic data** | | | | | |
| General Household Surveys, Statistics South Africa  <https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/StatsSA/about> | Annual household Survey | UCT | living circumstances of South African households: education, health, social development, housing, acces to services and facilities, food security and agriculture. | Sample survey data, uits are households and individuals | Research Project 2: provides information on socio-economic circumstances of residents within the Gauteng City-Region. |
| Quality of Life Surveys, Gauteng City-Region Observatory (GCRO)  <https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/GCRO/about> | Household Survey | UCT | quality of life, socio-economic circumstances, attitudes to service delivery, psycho-social attitudes, value-base and other characteristics of residents of the Gauteng City-Region. | Sample survey data, uits are households and individuals | Research Project 2: provides information on socio-economic circumstances and attitudes of residents within the Gauteng City-Region. |
| Global population (SEDAC) - Gridded Population of the  World (GPW), v4  https://sedac.ciesin. columbia.edu/data/c ollection/gpw-v4 | Distribution of human  population (counts and  densities) on continuous  global raster surface. Input data are extrapolated to  produce population estimates for 5-year intervals | PAIRS | Population counts and density estimates | Spatial: Global  coverage, 1km  grid res  Temporal: 5-  yearly; coverage  from Jan 2000 to Jan 2020 | Accounting for the  population exposed |
| News coverage  (https://www.gdeltpr oject.org/) | GDELT; Portion of news  coverage about specific area and time related to Covid-19. | PAIRS | Global events derived from worldwide news coverage. | Spatial: Global  coverage,  0.008192 deg. res | Example for production of spatial data layer for news events |

# Annex 2: Personal information processing agreement

The following agreement will be signed by each person (Operator under POPIA definitions) involved in processing personal information used by the project.

[Full name] hereby agrees to comply with the requirements of the POPIA Act of South Africa as regards the processing of personal information. These requirements include:

1. Only processing personal information for the purposes described in the HE²AT center research protocols
2. Only processing personal information that is required for these purposes
3. Not enabling or allowing access to personal information to anyone who does not have authorization for such access
4. Notifying the HE²AT, Steering Committee as the responsible party, if there is any reason to believe that personal information has been accessed or made available to an unauthorised person

Signed \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ at \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ on this \_\_\_ day of \_\_ in the year \_\_\_\_\_\_\_

# Annex 3: Ethics notification letter

A logo of a university

Description automatically generated

**APPLICATION FOR A NOTIFICATION REGARDING AN APPROVED STUDY**

|  |  |
| --- | --- |
| **PART 1: ADMINISTRATIVE**  *(Blocks will expand to contain the information required, no extra references or pages should be added)* | |
| Ethics Reference Number: |  |
| Study Title: |  |
| Phase of trial: |  |
| Protocol/Project/Study Number: |  |
| Approved Version/No. and Date: |  |
| Amended Version/No. and Date: |  |
| Health product being studied: |  |
| Sponsor/Funder/Donor: |  |
| Applicant: |  |
| Contact Person: |  |
| Address: |  |
| Cell No.: |  |
| E-mail address: |  |
| Date of Application: |  |

|  |
| --- |
| **PART 2: DETAILS OF NOTIFICATION** |

**Briefly provide:**

1. Motivation / Background:

2. Study Plan:

|  |  |
| --- | --- |
| I, the undersigned, agree to conduct/manage the above-mentioned study under the conditions as stated in this application | |
| Applicant/Principal Investigator:  Signature:  ………………………………………………… | Date  ……………………………………. |

1. Lakhoo, D.P., et al., *Protocol of an individual participant data meta-analysis to quantify the impact of high ambient temperatures on maternal and child health in Africa (HE(2)AT IPD).* BMJ Open, 2024. **14**(1): p. e077768.

2. Jack, C., et al., *Leveraging data science and machine learning for urban climate adaptation in two major African cities: a HE(2)AT Center study protocol.* BMJ Open, 2024. **14**(6): p. e077529.

1. https://www.gov.za/sites/default/files/gcis\_document/201409/3706726-11act4of2013protectionofpersonalinforcorrect.pdf [↑](#footnote-ref-1)
2. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [↑](#footnote-ref-2)
3. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [↑](#footnote-ref-3)