|  |
| --- |
|  |
| HE²AT Center Data Management Plan |
| Document version number 3.0 |

|  |
| --- |
| Developed by Christopher Jack  8-14-2024 |

## 

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](#_Toc174546606)

[Acronyms 4](#_Toc174546607)

[Definitions 5](#_Toc174546608)

[Stakeholders 8](#_Toc174546609)

[1.Scope 9](#_Toc174546610)

[2.Target Audience 9](#_Toc174546611)

[2. Purpose 10](#_Toc174546612)

[3. Background 12](#_Toc174546613)

[4. Data categories 14](#_Toc174546614)

[4.1 Health-related data 14](#_Toc174546615)

[Original study Data: 14](#_Toc174546616)

[Consortium Shared Data: 15](#_Toc174546617)

[RP1/RP2 De-Identified Data: 15](#_Toc174546618)

[Inferential Data: 15](#_Toc174546619)

[1.2. Climate/weather data 16](#_Toc174546620)

[1.3. Remote sensing data 17](#_Toc174546621)

[1.4. Areal/Geospatial socio-economic data 17](#_Toc174546622)

[5. Data management workflow 19](#_Toc174546623)

[5.1. Setting Up the Data Transfer Agreement 19](#_Toc174546624)

[5.2. Ethics NOTIFICATION OF New Databases 20](#_Toc174546625)

[5.3. Data Encryption and Transfer 21](#_Toc174546626)

[5.4. Data Storage and Encryption 22](#_Toc174546627)

[5.5. Data Indexing 22](#_Toc174546628)

[5.6. De-identification 23](#_Toc174546629)

[Geographic aggregation: 23](#_Toc174546630)

[Location jittering: 24](#_Toc174546631)

[5.7. Codebook remapping and harmonisation OF HEALTH DATA 26](#_Toc174546632)

[5.7.1 Pre-Processing: 26](#_Toc174546633)

[5.7.2 26](#_Toc174546634)

[Variable mapping: 26](#_Toc174546635)

[Mapping validation: 27](#_Toc174546636)

[Database population: 27](#_Toc174546637)

[Final data cleaning and analysis dataset creation: 27](#_Toc174546638)

[5.8. Data Integration 27](#_Toc174546639)

[5.9. Data analysis platforms 28](#_Toc174546640)

[Jupyter Hub 28](#_Toc174546641)

[IBM data analysis platforms 29](#_Toc174546642)

[Integration approaches 29](#_Toc174546643)

**[6.](#_Toc174546644)** [Data management, documentation and curation 30](#_Toc174546644)

**[7.](#_Toc174546645)** [POPIA compliance and protection of personal information 31](#_Toc174546645)

**[8.](#_Toc174546646)** [Data encryption 32](#_Toc174546646)

**[9.](#_Toc174546647)** [Network firewall and Virtual Private Network 33](#_Toc174546647)

**[10.](#_Toc174546648)** [Local authentication and authorization 33](#_Toc174546648)

**[11.](#_Toc174546649)** [Data sharing and open access 34](#_Toc174546649)

**[12.](#_Toc174546650)** [Procedure for making data available to bona fide researchers 35](#_Toc174546650)

**[13.](#_Toc174546651)** [Governance of data sharing 36](#_Toc174546651)

**[14.](#_Toc174546652)** [Data retention 38](#_Toc174546652)

**[15.](#_Toc174546653)** [Roles and responsibilities 39](#_Toc174546653)

**[16.](#_Toc174546654)** [Assessment and revision 41](#_Toc174546654)

[16.1. Assessment scope 41](#_Toc174546655)

[16.2. Revision process 41](#_Toc174546656)

**[Annex 1: Key data sources](#_Toc174546657)** [1](#_Toc174546657)

**[Annex 2: Personal information processing agreement](#_Toc174546658)** [1](#_Toc174546658)

**[Annex 3: Ethics notification letter](#_Toc174546659)** [2](#_Toc174546659)

**[Annex 4: Data Access Committee: Data Request form](#_Toc174546660)** [4](#_Toc174546660)

[References 4](#_Toc174546661)

# 

## Acronyms

|  |  |
| --- | --- |
| Acronym | Description |
| AOT | Aerosol Optical Thickness |
| BMGFKi | Bill and Melinda Gates Foundation Ki repository |
| CSAG | Climate System Analysis Group |
| DAC | Data Access Committee |
| DAP | Data Analysis Platform |
| DMAC | Data Management and Analysis Core of the HEÂ²AT Center |
| DMP | Data Management Plan |
| DS-I Africa | NIH Data Science Initiative Africa |
| DTA | Data Transfer Agreement |
| DUOS | Data Use Oversight System |
| ELSI | Ethical Legal and Social Implications Projects of DS-I Africa |
| FAIR | Findable, Accessible, Interoperable, and Reusable |
| FIPS | Federal Information Processing Standards |
| GCRO | Gauteng City-Region Observatory |
| HPC | High-Performance Computing |
| HSS | US Department of Human and Health Services |
| LDAP | Lightweight Directory Access Protocol |
| NDVI | Normalized Difference Vegetation Index |
| NIH | US National Institute of Health |
| NIR | Near-Infrared |
| NIST | National Institute of Standards and Technology |
| ODSP | Open Data Science Platform |
| PI | Principal Investigator |
| QoS | Quality of Service |
| RP1 | Research Project 1 of the HEÂ²AT Center |
| RP2 | Research Project 2 of the HEÂ²AT Center |
| Research Hubs | One of the seven NIH DS-I Africa Research hubs |
| SC | HEÂ²AT Center Steering Committee |
| SRTM | Shuttle Radar Topography Mission |
| TEC | Training and Engagement Core of the HEÂ²AT Center |
| TLS | Transport Layer Security |
| UCT | University of Cape Town |
| WWARN | Worldwide Antimalarial Resistance Network |
| sSA | sub-Saharan Africa |

|  |  |
| --- | --- |
| Term | Definition |
| Areal/Geospatial Socio-Economic Data | Represents socio-economic conditions such as household economic status, access to services, and dwelling type. Sourced from national census data and focused household and demographic surveys. |
| Bona Fide Researcher | An individual or entity engaged in legitimate scientific research with the objective of advancing knowledge in health data science, operating within the ethical, legal, and professional frameworks of academic and scientific research. |
| CSAG GitLab | The version control and collaboration platform used by the Climate System Analysis Group (CSAG) at the University of Cape Town (UCT) for managing the HEÂ²AT Center's data processes, documenting Data Reference Syntax (DRS), and storing data management code. It ensures transparent, version-controlled data handling accessible to authorized team members. |
| Climate/Weather Data | Includes observational-based datasets such as weather station observations, satellite proxy observations, and gridded climate data produced from atmospheric re-analysis and climate simulations. |
| Consortium Shared Data | Data that has undergone initial harmonisation and is shared among HEÂ²AT Center partners. This data is retained indefinitely or for five years post-project or DTA termination. |
| Controlled Access | A mechanism by which potentially personally identifiable research data is distributed to bona fide researchers upon submission of a data access request and supporting documentation. |
| Data Access Committee (DAC) | A committee responsible for reviewing and approving data access requests, and ensuring adherence to ethical, legal, and scientific standards. |
| Data Analysis Platform (DAP) | The platform used for conducting data analysis, typically including tools like JupyterHub. |
| Data Downloads | A modality of data sharing where researchers can download datasets to their local computing environments directly. |
| Data Management Plan (DMP) | A document outlining the procedures and standards for data acquisition, transfer, processing, storage, and access within the project. |
| Data Management and Analysis Core (DMAC) | The core component of the HEÂ²AT Center responsible for overseeing data management and analysis activities. |
| Data Protection Legislation | Any data protection or data privacy laws as may be applicable including but not limited to POPIA, the Electronic Communications and Transactions Act 26 of 2005, the Consumer Protection Act 68 of 2008, and the General Data Protection Regulation (GDPR). |
| Data Provider | The legal entity responsible for authorizing access to a dataset. |
| Data Subject | The individuals whose personal information is captured in health datasets. |
| Data Transfer Agreement (DTA) | A legal document outlining the terms and conditions under which data is shared between the provider and the recipient, addressing confidentiality, data use limitations, and compliance with relevant laws and guidelines. |
| Data Visiting | A modality of data sharing where data users access and analyse the data within the computing environment of the data provider. |
| Ethical Legal and Social Implications (ELSI) | Projects within DS-I Africa that focus on data science and research's ethical, legal, and social aspects. |
| Federal Information Processing Standards (FIPS) | Standards for federal computer systems that NIST develops. |
| Findable, Accessible, Interoperable, and Reusable (FAIR) | Principles that aim to improve the discovery, accessibility, interoperability, and reuse of data. |
| Gauteng City-Region Observatory (GCRO) | A research institute producing data on the quality of life in the Gauteng City-Region. |
| HE2AT Center Consortium | The consortium members jointly working on the HE2AT Project. |
| HE2AT Center Data Management Plan | The data management plan applicable to the RP1 Study as may be amended and updated from time to time. |
| HEÂ²AT Center | The Heat and Health in Africa Transdisciplinary Center, focused on heat-health research, capacity building, and engagement. |
| HEÂ²AT Center Steering Committee (SC) | The committee responsible for guiding the strategic direction and oversight of the HEÂ²AT Center. |
| Harmonization | Aligning various health datasets into a unified format. |
| Health-Related Data | Includes data collected from previous clinical cohort and trial studies, generally considered personal data due to its association with individual medical records and health events. |
| Integration | Aligning and integrating health and other various datasets into a unified format. |
| NIH Data Science Initiative (DS-I) Africa | An NIH initiative aimed at enhancing data science capacity and collaboration across Africa. |
| National Institute of Standards and Technology (NIST) | A US federal agency that develops and promotes measurement standards. |
| Near-Infrared (NIR) | A region of the infrared spectrum of light used in remote sensing applications. |
| Normalized Difference Vegetation Index (NDVI) | An index of plant 'greenness' or photosynthetic activity. |
| Open Access | Research data freely available in community-focused data repositories without major restrictions. |
| Open Data Science Platform (ODSP) | A platform facilitating the storage, retrieval, and processing of data for health research. |
| Operator | A person who processes Personal Data for a Responsible Party in terms of a contract or mandate without coming under the direct authority of that party. |
| Original Study Data | Raw, unprocessed data collected directly from various cohort studies and clinical trials. These data are owned by the data providers who conducted or commissioned the studies and is retained for five years following the termination of the project or the Data Transfer Agreement (DTA). |
| Personal Data | Any information relating to an identifiable living natural person and where it is applicable an identifiable existing juristic person. |
| Personally identifiable data | Data variables that enable the identification of an individual either directly through names, ID numbers, etc., or indirectly through combining other variables such as locations (GPS, street address), age, gender, and medical information. |
| Principal Investigator (PI) | The lead researcher responsible for the conduct of a research project. |
| Processing | Any operation or set of operations which is performed upon Personal Data whether or not by automatic means such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction. |
| Quality of Service (QoS) | A measure of the overall performance of a service such as a survey or data collection process. |
| RP1 De-identified Data | Data with the following information deleted |
| Re-analysis | A dynamical model simulation of historical climate evolution continuously nudged by observations to provide an approximate historical representation of the climate system. |
| Remote Sensing Data | Data derived from satellite sensors, including optical imagery and indicators of physical measures like land surface temperature, soil moisture estimates, and vegetation condition. |
| Research Hubs | Seven NIH DS-I Africa U54 grants that are called Research Hubs and who contribute to and utilize shared data resources. |
| Research Project 1 (RP1) | The first major research project of the HEÂ²AT Center, focusing on individual participant data meta-analysis of heat-health impacts. |
| Research Project 2 (RP2) | The second major research project of the HEÂ²AT Center, focusing on urban heat vulnerability and early warning systems. |
| Responsible Party | A public or private body or any other person which alone or in conjunction with others determines the purpose of and means for Processing Personal Data. |
| Sensitive data | Data that pertains to an individual's personal information, health, finances, occupation, etc. |
| Shuttle Radar Topography Mission (SRTM) | A mission that obtained elevation data for most of the Earth using radar interferometry. |
| Social Media Data | Data representing activity on social media platforms, initially Twitter, and includes data like Google Search trends and Facebook mobility data. Used for analysing public perceptions. |
| Training and Engagement Core (TEC) | The core component of the HEÂ²AT Center responsible for training and capacity-building activities. |
| Transport Layer Security (TLS) | A cryptographic protocol designed to provide secure communication over a computer network. |
| US National Institute of Health (NIH) | The primary agency of the United States government responsible for biomedical and public health research. |
| Worldwide Antimalarial Resistance Network (WWARN) | A network focused on monitoring and understanding antimalarial drug resistance. |
| sub-Saharan Africa (sSA) | A geographical region of Africa located south of the Sahara Desert. |

# 

* **CSAG/UCT Data Management Team:** The Climate System Analysis Group (CSAG) at the University of Cape Town (UCT) is responsible for managing data storage, harmonization, and integration processes within the HE²AT Center.
* **Climate Data Providers**: Organizations or agencies that supply data related to climate and weather, such as weather station observations, satellite data, and gridded climate data. Examples include national meteorological services and international climate data repositories like Copernicus Climate Data Store (CDS).
* **Compliance Officers:** Individuals responsible for monitoring and enforcing compliance with data security and privacy regulations, ensuring that all data management activities adhere to relevant laws and guidelines.
* **Core HE²AT Team:** A group of named key personnel responsible for the initial processing and de-identification of the Original Source Data.
* **Data Access Committee (DAC):** A committee that evaluates and approves requests for data access, ensuring that data sharing complies with ethical, legal, and scientific standards.
* **Data Providers:** Legal entities or organizations responsible for authorizing access to their datasets and ensuring that data sharing complies with legal and ethical standards.
* **DSI Partners:** Partners within the NIH Data Science Initiative (DS-I) Africa who contribute to and utilize the shared data resources, collaborating on data management practices to ensure alignment with the broader DS-I Africa objectives.
* **Ethics Committees**: Institutional bodies that review and approve the ethical aspects of research projects, ensuring that studies comply with ethical standards and protect participants' rights and welfare.
* **Harmonization Team Members**: Researchers and data scientists who work on aligning and integrating various datasets into a unified format, ensuring consistency and usability for analysis.
* **Health Data Providers**: Entities responsible for collecting and providing biomedical data, including clinical trial coordinators, cohort study administrators, hospitals, and research institutions involved in health-related studies.
* **Health Experts:** Specialists in health-related fields who validate the harmonization of biomedical data and ensure the accuracy and reliability of the integrated datasets.
* **IBM Data Analysis Platforms Team:** The team responsible for managing IBM's suite of data analysis tools and platforms, such as IBM Watson Studio, which support advanced analytics and machine learning within the HE²AT Center.
* **OpenAI:** An artificial intelligence research organization that provides large language models (LLMs) to assist with variable mapping suggestions, data transformation recommendations, and other AI-driven insights.
* **Partner Researchers:** External researchers and scientists collaborating with the HE²AT Center and utilizing the integrated datasets for various research projects and analyses.
* **Post-Funding Data Management Team:** The team responsible for managing data after the funding period ends, ensuring the continued maintenance and availability of inferential data and transferring data to appropriate repositories as needed.
* **Socio-Economic Data Providers:** Organizations that provide data on socio-economic conditions, such as national census bureaus, household survey agencies, and specialized observatories like the Gauteng City-Region Observatory (GCRO).
* **Training Hubs:** Facilities that provide training and capacity-building activities for researchers, using shared data for educational purposes and contributing to the development of data management skills within the consortium.

# Scope

This Data Management Plan (DMP) applies to all data acquired and produced as part of the activities of the HE²AT Center project, including pilot projects.

# Target Audience

The scope encompasses various stakeholders, each with distinct roles and responsibilities as outlined below:

* **Data Providers**: These are the legal entities or organisations responsible for authorising access to their datasets. They ensure that data sharing complies with legal and ethical standards.
* **DSI Partners**: Partners within the NIH Data Science Initiative (DS-I) Africa contribute to and utilise the shared data resources. They collaborate on data management practices and ensure alignment with the broader DS-I Africa objectives.
* **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.
* **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).
* **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.

**Researchers Requests 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 transfer agreements, ethical guidelines, and any conditions set by the Data Access Committee (DAC).

# Purpose

The Data Management Plan (DMP) aims to establish comprehensive procedures and standards for transferring, processing, storing, and accessing data within the HE²AT Center project. This plan aims to ensure data integrity, security, and availability, while facilitating effective data sharing and collaboration among all stakeholders.

Specifically, the DMP seeks to:

* **Standardize Data Management Practices**: Define uniform procedures and standards to streamline data handling processes across the consortium, ensuring consistency and reliability.
* **Enhance Data Security and Privacy**: Implement robust measures to protect sensitive and personally identifiable data, complying with relevant legal and ethical standards.
* **Facilitate Data Sharing and Access**: Clear guidelines and DTAs promote transparency and collaboration, and support data sharing within the consortium, other DS-I Africa projects, and with external researchers.
* **Support Data-Driven Research**: Provide a framework that enables researchers to efficiently access and utilise data for innovative research, contributing to the HE²AT Center's mission of advancing heat-health knowledge and interventions.
* **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, data sharing outside the consortium is further supported by associated Data Transfer Agreements between the HEAT Center and external partners, ensuring that all data management activities, both inside and outside the Consortium, 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 by 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 nature of associations between exposure to high ambient temperatures and selected health outcomes in pregnant women and children within the first two years of life. Such techniques have not yet been employed in climate change and health. An IPD can overcome many limitations of traditional analyses of individual datasets and biases in classic systematic review methodology. The project has systematically identified potentially eligible African cohort studies or trials through a systematic mapping of studies on pregnant women and children in Africa. Data are being harmonised through re-coding of 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 growth in urban populations, geographical extent of cities and over-burdened health services in African cities, coupled with rising temperatures and Urban Heat Island phenomenon, pose significant public health challenges. High ambient temperatures cause considerable morbidity and mortality in urban areas, influenced by temperature gradients, socio-environmental factors and the characteristics of the built environment. This project, conducted in Abidjan, Ivory Coast, and Johannesburg, South Africa, will investigate heat exposure risks to develop an Early Warning System for vulnerable groups.

We will use data science methods, including natural language processing and predictive geospatial analysis, to integrate diverse data streams. Potential mediating and confounding factors, such as urban form, differentials in socio-economic status and vegetation indices will be included in exposure-response analyses involving high-resolution climate data and health outcomes. The project aims to estimate historical and future heat hazards and develop a predictive model linking heat exposure to health outcomes under a range of emission and development scenarios. Health data will come from cohort and clinical trials in the target cities.

Various dissemination methods for the early warning system will be explored, including a web-based application. The project aligns with DS-I Africa objectives, and may expand to include other Research Hubs and cities across Africa.

Footnote the ref: [2].

# Data categories

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

## Health-related data

**4.1.1 Health outcome data for RP1 and RP1**

These data will include data collected from previous clinical cohort and trial studies (e.g. maternal and child cohort studies, HIV treatment trials and COVID studies). Data sources and criteria for inclusion of studies vary between the two RPs.

**Health-related data for RP1** were identified through a systematic review, mapping large (n>1000) longitudinal studies on pregnant women and children 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 (See RP1 Protocol Document for full details).

**Health-related data 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 included 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 (See RP2 Protocol Document for full details).

**4.1.2 Categories of data, by harmonization and deidentification status**

Health data takes four forms, moving from original individual-level study data to aggregated inferential data. This categorization helps ensureence,

### Original study Data

This category includes raw, unprocessed health data collected directly through various cohort studies and clinical trials that fulfill the study eligibility criteria. 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. The source data is protected pseudonymous data, meaning it has been de-identified, but it still contains indirect identifiers such as dates and geolocation information. While these identifiers do not directly reveal personal identities, they could be used for identification if cross-referenced with other information.

### Consortium Shared Data

Once the source data undergoes harmonisation, it becomes consortium-shared data. The data is also integrated with climate and other data, as described below. Unless data providers state otherwise, the HE²AT Center and Data Access Committee now make decisions around data usage and access at this stage. This level of decision-making stems from the team having performed significant alterations, harmonization, and integration work on the data. However, original data providers may elect to stipulate different terms of data use and access, 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 HE2AT Center partners and can be extended to new partners who join the Consortium.

This data is considered protected pseudonymous data as per POPIA guidelines as several steps have been done to enhance privacy protections. For example, age data will be reported in years, rather than exact birthdates and geographic data will be jittered.

### RP1/RP2 De-Identified Data

Data that has been further processed and de-identified falls into this category. The HE²AT Center retains ownership of this de-identified data, which is stored indefinitely unless the Data Provider stipulates otherwise. 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. Age data will be reported in five-year age bands rather than an exact birthdate, and the geographic data will be reported in larger units, such as cities or districts.

### Inferential Data

The final category is inferential data, aggregated and synthetic data derived from the analysis of the preceding data categories. The HE²AT Center owns these details and retains them 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.

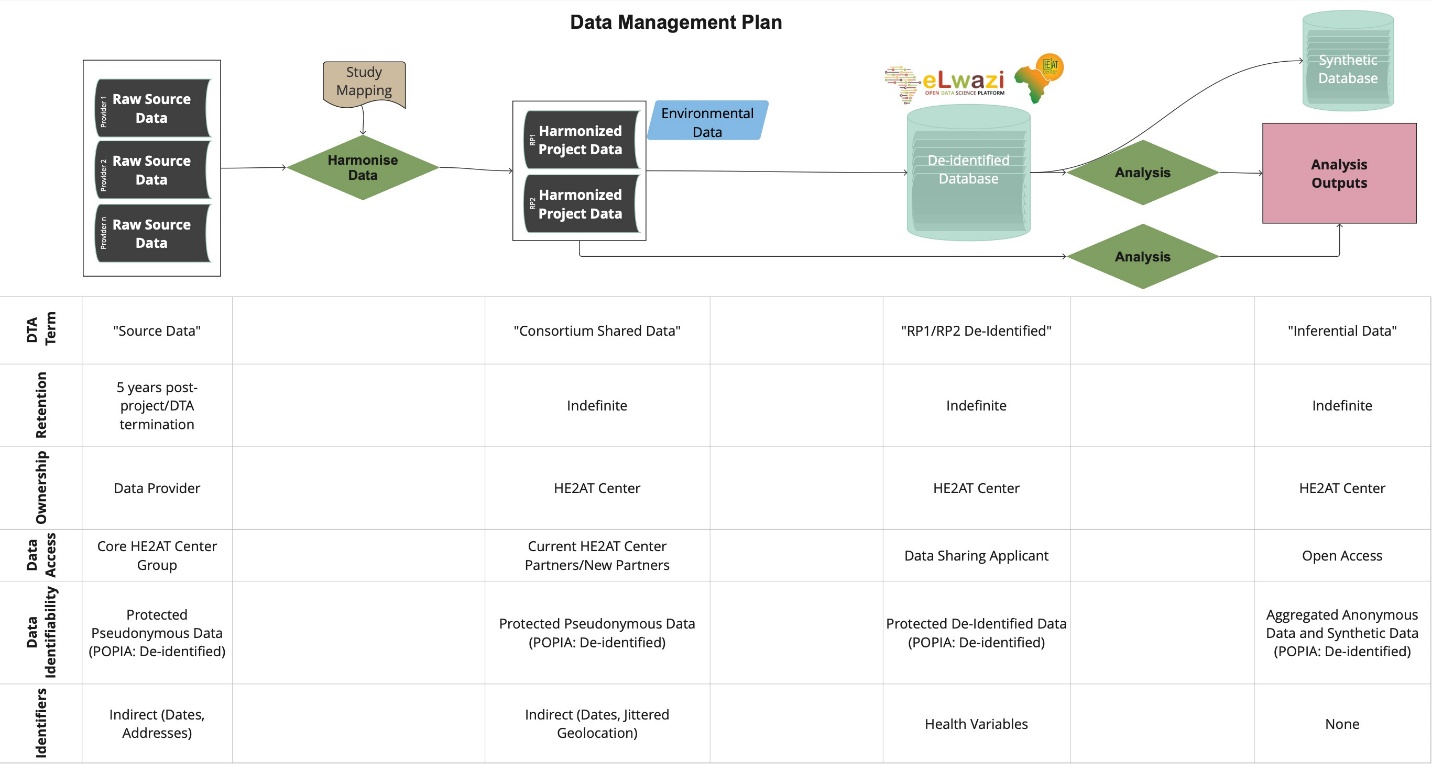


Figure 1:HE²AT Center Data Management Workflow

1.2. Climate/weather data

This data includes 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.

Obtaining climate-related data will mostly involve 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. CSAG/UCT will manage and update the primary data index to help make 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. We will cite the source(s) whenever the data are used. .

## 4.2. Remote sensing data

The 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 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.

While **climate/weather data** and **remote sensing data** may involve accessing similar open data repositories and storage systems, they serve different purposes and do not overlap in their data types. Climate/weather data primarily focuses on atmospheric conditions and forecasts, whereas remote sensing data pertains to physical and environmental measurements derived from satellite imagery. Both data types are managed and updated by CSAG/UCT to ensure accessibility and utility for research purposes.

## 4.3. Areal/Geospatial socio-economic data

These 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 open 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. Still, versions may already exist or can be uploaded on IBM’s system to enable analysis through IBM analytical systems.

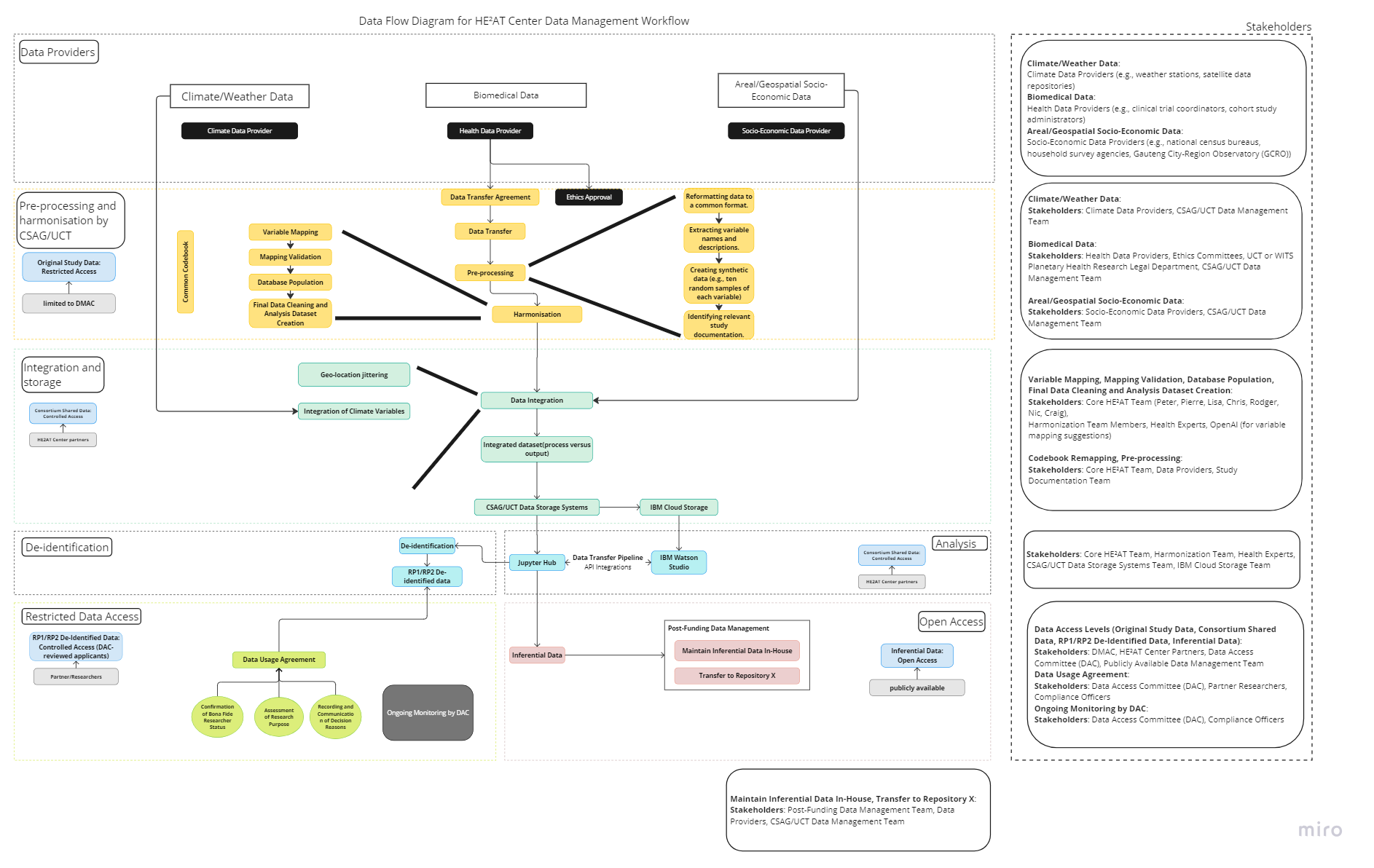
South African census data is already available through the UCT DataFirst data repository. GCRO Quality of Life Survey data is available through the GCRO open data platform, which directs queries with GCRO.

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.

# 

# 5. Data management workflow

Figure 2 below maps out the data processing workflow. Each element within this workflow is described below in detail.



***Figure 2: Detailed data flow diagram for HE²AT center data management workflow***

This data flow diagram illustrates the full data management workflow within the HE²AT Center. It encompasses the stages from data acquisition to data sharing and compliance. The diagram details the roles of data providers, the initial steps conducted by CSAG/UCT, data harmonisation, storage and indexing, integration and analysis, data access levels, and ongoing monitoring for compliance.

## 5.1. Setting Up the Data Transfer Agreement

A Data Transfer Agreement (DTA) is a crucial step in the health data transfer process. It is a formal agreement between each health data provider and UCT or WITS Planetary Health Research, enabling the transfer, processing, analysis, and 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.

## 5.2. Ethics NOTIFICATION OF New Databases

As new databases become available and DTAs are completed, the following steps will be taken to notify the Wits Health Research Ethics Committee (Medical). These processes aim to secure ethics oversight, and maintain the integrity and compliance of the HE²AT Centre's research activities. The HE2AT Center will provide additional information on the individual 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 from Johannesburg**

1. **Data Transfer Agreements (DTAs):**
   * Prior to notification, the DTAs must be signed by the data provider and the legal representative of UCT or Wits legal office.
   * Each DTA will be submitted with the protocol as part of the ethics notification process.
2. **Ethics Approvals:**
   * The Wits Human Research Ethics committee will review the period since the study closure to identify any potential concerns.
   * Each study will be individually assessed to ensure ethical standards have not changed since the original approval.
   * Original ethics approvals for the studies remain valid as only secondary data analyses areas are being done in a study (RP2) that has ethics approval.
3. **Notification Form:**
   * A two-page ethics notification form will be completed and submitted to the ethics committee (Appendix 3). It includes the study background, motivation, design, and the data fields to be collected.
   * The notification form must be submitted along with fully signed DTAs.
   * Wits HREC (Medical) will review the notifications and identify any concerns for discussion.

Data transfer and use may proceed once the ethics committee signs the notification and returns it to the HE2AT Center.

## 5.3. Data Encryption and Transfer

Once the Data Transfer Agreement has been agreed upon and signed, and, in the case of RP2, the ethics committee has raised no ethical concerns, the data provider will initiate the transfer of data to the UCT data platform. The detailed steps for data encryption and transfer are as follows:

**Encryption of Personally Identifiable Data**

1. **Encryption by Data Provider**:
   * **Encryption Process**: If the data contains personally identifiable information (PII), the data provider will encrypt the data before transfer. This involves converting the data into a coded format that is unreadable without a decryption key, using strong encryption algorithms such as AES-256 (Advanced Encryption Standard with a 256-bit key).
   * **Encryption Standards**: The AES-256 standard, established by the US National Institute of Standards and Technology (NIST), ensures a high level of security, making it extremely difficult for unauthorized parties to decrypt the data without the appropriate key.
2. **Secure Transfer Protocols**:
   * **Transport Layer Security (TLS)**: The encrypted data will be transferred using a secure service that employs encrypted data transport protocols like TLS. TLS ensures that the data remains encrypted during transmission, protecting it from interception or tampering.
   * **FileSender Example**: An example of a secure transfer service is FileSender. FileSender utilises TLS to encrypt the data during transfer, ensuring that the data remains secure from the point of departure to its destination. The service also includes features such as user authentication and detailed logging of transfer activities.
3. **Transfer Execution**:
   * **Initiating Transfer**: The data provider will initiate the transfer by uploading the encrypted data to the chosen secure transfer service (e.g., FileSender).
   * **Transfer Monitoring**: The transfer process will be monitored in real-time to ensure it is completed successfully without interruptions or errors. The data provider and the recipient (UCT data platform) will receive notifications upon successfully completing the transfer.
4. **Data Receipt and Verification**:
   * **Data Receipt**: The UCT data platform will receive the encrypted data file upon successful transfer.
   * **Verification Process**: The recipient team at UCT will verify the data's integrity and completeness. This includes checking the file size, verifying checksum, and confirming the encryption has not been compromised.
   * **Decryption Process**: Once verification is complete, authorised personnel at UCT will decrypt the data using the provided decryption key. The decrypted data will then be stored securely on the UCT data platform, ready for further processing and analysis.

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

## 5.4. Data Storage and Encryption

Once the source dataset has been transferred to UCT, if it constitutes personally identifiable data or if stipulated by the DTA, the dataset will be encrypted for storage. The detailed process is as follows:

1. **Encryption for Storage**:
   * **AES-256 Encryption**: The dataset will be encrypted using 256-bit AES (Advanced Encryption Standard), established by the US NIST. This ensures that the data is protected with a high level of security.
   * **Access Control**: Encryption keys will be available only to the minimum number of people (the Core HE²AT Center Group) required to implement any anonymization or data minimization processes. This group includes key personnel responsible for data management and security who have completed the Personal information processing agreement (Annex 2).
2. **Metadata Management**:
   * **Separate Storage for Metadata**: Metadata, including dataset descriptions, descriptions of the original protocols, and codebooks, will be decrypted and stored separately from the encrypted dataset. This allows for efficient metadata indexing and software code development by a larger group of HEAT Center staff, while maintaining the security of the main dataset.

## 5.5. Data Indexing

Data available on the CSAG/UCT data storage system and IBM’s storage system will be indexed using an appropriate metadata standard. The CSAG/UCT data platform will be used to make the index available to the Consortium partners and bona fide researchers beyond the Consortium. The indexing process includes:

1. **Metadata Standards**:
   * **eLwazi Integration**: Data sharing outside of HE²AT Data will utilize the eLwazi platform, ensuring compliance with broader data-sharing protocols.
   * **Data Reference Syntax (DRS)**: CSAG implements a Data Reference Syntax (DRS), a structured mapping from a controlled vocabulary of metadata elements to a directory and file naming syntax. This is standard practice within climate data management and will be continued for climate and remote sensing datasets.
   * **Health Data Indexing**: Health data will be indexed using a codebook that applies relevant ontologies, ensuring consistency and discoverability.
2. **Documentation and Integration**:
   * **CSAG GitLab Wiki**: The current DRS is documented on the CSAG GitLab Wiki, providing a reference for the indexing process.
   * **DSI-Africa Open Data Science Platform (ODSP)**: Integration with the ODSP metadata index will ensure that metadata propagates to the ODSP system, making datasets discoverable through ODSP metadata queries.

By following these detailed procedures, the HE²AT Center ensures that data encryption, transfer, storage, and indexing are conducted in a secure, efficient, and compliant manner, aligning with the highest standards of data management and protection.

## 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[3].[[1]](#footnote-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.

### 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.

A diagram of different colors of a circle

Description automatically generated

Figure 2: 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.

**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.

**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.[[2]](#footnote-2)

## Codebook remapping and harmonisation OF HEALTH DATA

Data harmonisation is crucial for integrating diverse HEALTH datasets into a unified format, enabling comprehensive analyses across disparate data sources. The workflow from Original Source data to de-identified data is as follows:

### 5.7.1 Pre-Processing:

This initial step is performed by the 'core' HE²AT team (members with access to 'incoming' data). It involves:

* Reformatting data to a common format.
* Extracting variable names and descriptions.
* Creating synthetic data (e.g., ten random samples of each variable).
* Identifying relevant study documentation.

This data is periodically backed up to a secure and encrypted disk, separate from the wider HPC system.

### **5.7.2** Variable mapping:

This step involves the Core group and a larger set of HEAT Center staff who utilise synthetic data, variable metadata, and study documentation created in the pre-processing step. Data is shared with OpenAI to leverage LLM description completions, variable mapping suggestions, and data transformation recommendations. Harmonisation team members complete a form to map source variables to a standard set of ontologies and recode variables to a standard data model.

### Mapping validation:

The 'core' HE²AT team validates the variable mappings created in the previous step using the reformatted source data. This validation is conducted first by a health expert and then by the core harmonising team. The mappings and transformation instructions from this step are version-controlled using CSAG GitLab.

### Database population:

Using the mappings and transformation instructions from the previous steps, the source data is transformed and used to populate a final database. This step is managed by the 'core' team, and the final database is made available to the full consortium. A second anonymisation of patient identifiers occurs at this stage, and locations below the ward level (e.g., street addresses) are excluded.

### Final data cleaning and analysis dataset creation:

A final data cleaning step ensures values fall within expected ranges, provisions for identifying twins and multiple pregnancies are made, and duplication errors are identified. A sanity check ensures data integrity. Linking geospatial data is an additional step performed during the analysis stage.

## Data Integration

Integrated datasets, along with associated documentation provided through the GitLab platform, will be available for analysis via the Jupyter Hub platform to partner researchers. Depending on the conditions of the original Data Transfer Agreements (DTAs), partner researchers may need to agree to specific data usage agreements.

The complete health data processing workflow is mapped in Figure 2 below. Key steps include:

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 pose a re-identification risk. For example, the 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 research project, translating variables into the target units, converting real values into categories, and translating between categorization schemes. This step also involves recording geographic information, such as:
   * **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.

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

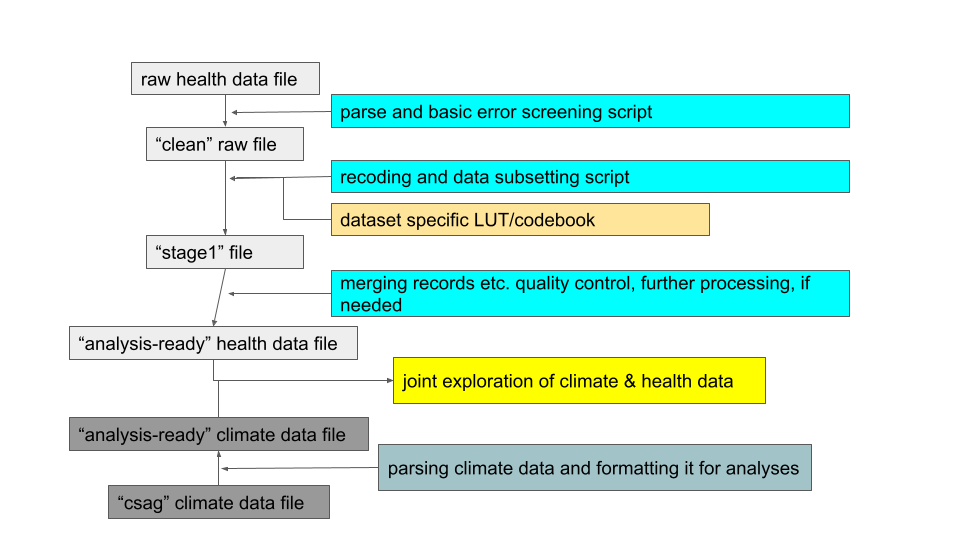


Figure 2: Detailed data processing for health data ingestion

## Data analysis platforms

Data analysis will be facilitated through the CSAG/UCT Jupyter Hub and IBM platforms, providing robust and scalable environments for processing and analysing the HE²AT Center datasets.

### Jupyter Hub

Jupyter Hub is a collaborative, web-based Python coding environment that allows analysts to develop and execute analysis code using a browser interface. Key technical details include:

* **Web-Based Platform**: Researchers can access Jupyter Hub through a web browser, providing a user-friendly interface for coding and data analysis.
* **Python Environment**: Supports the development of analysis code in Python, leveraging a wide range of libraries and frameworks for data science, machine learning, and statistical analysis.
* **CSAG High-Performance Computing (HPC) Integration**: Analysis code executed in Jupyter Hub runs on the CSAG HPC platform, providing high computational power for processing large datasets.
* **Data Accessibility**: Datasets stored on CSAG storage servers are directly accessible within the Jupyter Hub environment, allowing seamless data retrieval and manipulation.
* **Collaboration**: Jupyter Hub supports collaborative work, enabling multiple analysts to share and work on the same notebooks, fostering teamwork and knowledge sharing.

### IBM data analysis platforms

IBM offers a suite of interfaces and tools for rapid and complex data analysis, which will be available to HE²AT Center analysts upon request. These include:

* **IBM Watson Studio**: A comprehensive data science platform that provides tools for data preparation, modelling, and deployment. Analysts can build and train machine learning models, perform data exploration, and create visualisations.
* **IBM Cloud Pak for Data**: An integrated data and AI platform that allows analysts to collect, organise, and analyse data at scale. It includes various services for data governance, data science, and AI model management.
* **IBM SPSS Modeler**: A powerful tool for building predictive models using a visual interface. It supports various statistical and machine learning techniques and integrates with other IBM data platforms.
* **IBM Analytics Engine**: Provides a scalable Hadoop and Spark environment for processing large datasets and running advanced analytics.

### Integration approaches

To enhance the analytical capabilities, approaches to integrate data from IBM into the CSAG/UCT Jupyter Hub will be developed. This integration will allow analysts to leverage the strengths of both platforms:

* **Data Transfer Pipelines**: Establishing automated pipelines to securely transfer data between IBM and CSAG/UCT storage systems, ensuring data consistency and availability.
* **API Integrations**: APIs are utilized to enable seamless data access and manipulation across platforms, allowing Jupyter Hub notebooks to fetch and process data stored in IBM systems.
* **Unified Access**: Developing interfaces within Jupyter Hub that allow analysts to directly invoke IBM services, such as running IBM Watson Studio experiments or accessing IBM Cloud Pak for Data resources.
* **Security and Compliance**: Ensuring that all data transfers and integrations comply with security standards and data governance policies, maintaining the confidentiality and integrity of the data.

By leveraging these platforms and integration approaches, the HE²AT Center will provide a comprehensive and flexible environment for data analysis, enabling advanced research and insights into heat and health impacts.

# 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 procedures 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 program 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, guided by relevant legislation specific to each dataset, including country-specific laws on personal/sensitive data and cross-border data transfer. The development and negotiation of these data-sharing agreements are managed by the DMAC in conjunction with research projects, as they are the primary interface with the data sources.

The Protection of Personal Information Act (POPIA) of South Africa (2013) sets limitations on processing personal information but provides a legal basis for its use in scientific research. POPIA works alongside other South African legislation, such as the Constitution, the National Health Act No. 61 of 2003, and the Department of Health guidelines on Ethics in Health Research (2015). The law offering the most comprehensive protection for individuals' rights takes precedence.

Section 6 of POPIA states that the Act does not apply if personal information has been de-identified to the extent that re-identification is virtually impossible. Many health databases for HE²AT will meet this criterion. For those that do not, the following POPIA sections provide a basis for processing health data:

* **Section 15(1)**: Personal information processing must align with the purpose of collection.
* **Section 15(3)(e)**: Allows processing for historical, statistical, and research purposes, regardless of the original purpose of collection. This is crucial for HE²AT, as health datasets were collected before the project began.
* **Section 18(1)**: Requires informing data subjects about the processing of their personal information.
* **Section 18(4)(f)**: Provides an exemption for informing data subjects if the information is used for historical, statistical, or research purposes.
* **Section 14(2)**: Allows retention of personal information for research as long as safeguards prevent its use for other purposes.
* **Section 16**: Mandates reasonable measures to maintain data quality.
* **Section 17**: Requires clear documentation of all processing activities.
* **Section 19**: Requires security measures to prevent unlawful access to or processing of personal information.
* **Section 20**: Specifies requirements for individuals processing personal information. We maintain a continually updated list of authorized personnel with restricted access to personal information through passwords and other security measures.
* **Section 21**: Requires a written contract between the responsible party and operators implementing processing. This contract, presented in Annex 1, mandates that operators inform the responsible party if unauthorized access to personal information is suspected.

# Data encryption

To ensure the highest level of data security, primary data (before processing) will be stored in encrypted form using 256-bit Advanced Encryption Standard (AES), a cryptographic standard established by the US National Institute of Standards and Technology (NIST). The detailed encryption process includes:

**Encryption Standards and Protocols**

* **AES-256 Encryption**: This encryption method utilises a 256-bit key, making it highly secure and resistant to brute-force attacks. AES-256 is widely recognised for its robustness and is used globally to protect sensitive data.
* **NIST FIPS 140-2 Compliance**: The United States government requires the use of cryptographic modules validated to NIST Federal Information Processing Standards (FIPS) 140-2 for encrypting all data classified as Sensitive but Unclassified (SBU) or above. This ensures that the encryption methods meet stringent security requirements.

**Access Control**

* **Encryption Key Management**: Encryption keys are only available to a minimum number of authorised personnel necessary to perform de-identification processing. This minimises the risk of unauthorised access to the encryption keys and, by extension, the encrypted data.

**Secure Data Transfer**

* **TLS Encryption**: For datasets that include personal identifiers (i.e., personally identifiable data), the data will be encrypted by the original data holder and transferred to UCT using Transport Layer Security (TLS). TLS provides secure data transmission over the internet, preventing interception and unauthorised access during the transfer process.

**Storage Isolation**

Encrypted data will be stored in an isolated virtual storage server, ensuring enhanced security and controlled access. Key details include:

* **Isolated Storage Servers**: The encrypted data is stored on dedicated virtual servers isolated from other systems. This isolation reduces the risk of data breaches by restricting access to the servers where the data resides.
* **Restricted Access**: Only the private data team, a select group of authorized personnel, can access these isolated storage servers. This team is responsible for managing and maintaining the security of the stored data.

# Network firewall and Virtual Private Network

The CSAG compute infrastructure is integrated into the UCT intranet, benefiting from UCT’s comprehensive security policies and services. Key network security measures include:

* **Firewall Protection**: UCT implements a Cisco firewall as a barrier between the internal network and external threats, ensuring that only authorised access is allowed. The UCT firewall policy outlines the specific configurations and rules applied to protect the network.
  + [UCT Perimeter Firewall Policy](http://www.icts.uct.ac.za/uct-perimeter-firewall-policy)
* **Virtual Private Network (VPN)**: UCT employs a Cisco VPN service, allowing authorised users to access the UCT intranet from remote locations securely. The VPN encrypts all traffic, preventing unauthorised access and maintaining data confidentiality.
* **Access Control**: Network access to CSAG servers and computer services is carefully controlled and limited to authorised users. This ensures that only individuals with proper credentials can access sensitive data and systems.

# Local authentication and authorization

In addition to the broader UCT security measures, the CSAG/UCT platform employs its own authentication and authorisation protocols to secure access to restricted datasets further:

* **Independent Authentication Service**: The CSAG/UCT platform uses a Linux filesystem and Lightweight Directory Access Protocol (LDAP) for authentication. This service verifies the identity of users attempting to access the system.
* **Authorization Controls**: Access to restricted datasets is controlled through both UCT authentication protocols and internal CSAG Data Management Plan (DMP) authentication and authorisation mechanisms. This dual-layer security approach ensures that only authorised personnel can access sensitive data, providing an additional layer of protection.
* **Compliance with UCT Security Policies**: All authentication and authorisation activities are conducted in compliance with UCT's comprehensive information security policies.
  + [UCT Information Security Policy](http://www.icts.uct.ac.za/sites/default/files/image_tool/images/286/UCT_Information_Security_Policy_PC03_2020.pdf)

By implementing these detailed encryption, storage isolation, network security, and local authentication measures, the HE²AT Center ensures the protection and confidentiality of sensitive data throughout its lifecycle.

# 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 Transfer 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 bona fide researchers

The procedures outlined in the DSI-Africa Data Sharing Guideline govern the process of making data available to qualified individuals for Research Project 1 (RP1) and Research Project 2 (RP2). This ensures compliance with ethical, legal, and scientific standards, maintaining the integrity and security of the data while facilitating its responsible use.

**Step 1: Development of Data Access Request Form** The DAC will develop a comprehensive Data Access Request Form for researchers outside the HE²AT Center team. This form must be completed by individuals requesting access to the RP1/RP2 de-identified data and include:

* **Applicant Information**: Details about the applicant and their institution, including name, address, and the nature of its research activities.
* **Consortium Membership Status**: Indication of whether the applicant is a member of the DSI-Africa Consortium. Non-members must provide detailed information about their organisational affiliation.
* **Dataset Identification**: Specific datasets to which access is sought.
* **Research Purpose**: A proposal outlining the intended research, including the research objectives and significance.
* **Data Sharing Modality Preference**: Specification of whether data access is sought through download or data visiting, with a justification based on the research and data sensitivity.
* **Ethics Approval**: Proof of ethics approval from the applicant's institutional ethics committee.
* **Data Protection Measures**: Detailed description of statutory, organisational, and technical measures in place at the receiving institution to safeguard the confidentiality and safety of the data.

**Step 2: Preliminary Screening** Upon submission, the Data Access Request Form undergoes a preliminary screening by the Data Access Committee (DAC) to confirm completeness and basic compliance with the Consortium’s requirements.

**Step 3: Review by Data Access Committee** The DAC will evaluate the request based on the following criteria:

* **Alignment with Data Use Permissions**: The request must align with the data use permissions outlined in the Data Transfer Agreement (DTA).
* **Privacy Risk Assessment**: Potential risks to the privacy of research participants must be identified and addressed satisfactorily.

**Non-Overlap with Ongoing Research**: The proposed research should not overlap with or pre-empt the results of ongoing Consortium projects. If overlap is identified, a temporary embargo may be imposed.

**Step 4: Recording and Communication of Decision Reasons**

Once a decision is made, the DAC records the reasons for its approval, conditional approval, or denial. This documentation ensures transparency and provides valuable feedback to applicants. The decision and any conditions attached are communicated to the applicant clearly and promptly. Approved requests will include detailed instructions for accessing the data and any requirements or conditions the applicant must meet.

**Step 5: Data Transfer and Encryption**

Upon approval and fulfillment of all conditions, including ethics approval, the data will be transferred. If the data contains personally identifiable information, it will be encrypted by the data provider before transfer using a secure service employing encrypted data transport protocols such as TLS (Transport Layer Security). Examples of such services include FileSender and Box.com.

**Step 6: Ongoing Monitoring** The DAC will continue to monitor compliance with the terms of the Data Transfer Agreement, including periodic reviews and audits if necessary.

# Governance of data sharing

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

**Protection Measures**

**Data Protection Impact Assessment (DPIA)**: Not all applications require a DPIA. For example, de-identified data that cannot be re-identified does not require a DPIA. However, if personal information is present, a DPIA is mandatory. This assessment involves identifying privacy risks and implementing mitigation strategies such as data encryption and reducing the amount of personal information shared.

Recipient shall develop and maintain a data protection impact assessment regarding the Processing of Personal Data under this Agreement. Data Provider shall cooperate with and assist Data Recipient in the development of the data protection impact assessment and/or with prior consultations with government authorities that may be required.

**Data Transfer Agreement (DTA)**: A DTA must be executed between the data provider and recipient before any data is shared. The DTA outlines the terms and conditions for data use, ensuring compliance with legal and ethical standards. It includes a prohibition on on-sharing data, ensuring the recipient does not pass it to third parties.

**Post-Funding Data Management**

The DSI Africa Consortium provides two options for sustainable data management post-funding:

**Maintain Inferential Data In-House**: The main host institution continues to store, preserve, and secure Inferential Data, managing access requests and maintaining a data visiting infrastructure if applicable.

**Transfer to Repository X**: Inferential Data is transferred to Repository X, managed by an NIH-appointed DAC, ensuring long-term preservation and accessibility.

Regardless of the chosen option, data must not be deleted, ensuring its availability for future scientific inquiry.

**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:**

**Step 1: Confirmation of Bona Fide Researcher Status** External applicants, i.e., those not affiliated with the HE²AT Center Consortium, undergo a thorough verification process to confirm their status as bona fide researchers. The DAC reviews the provided credentials, affiliations, and 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.

**Step 2: Assessment of Research Purpose** The DAC applies three criteria for access:

The data request must be aligned with the Data Use Permission.

Potential risks to the privacy of research participants associated with the proposed use of the data, where relevant, have been identified and addressed satisfactorily.

The proposed research does not overlap with or pre-empt the results of ongoing Consortium projects. If overlap is identified, the DAC may impose a temporary embargo on data sharing rather than declining access. This 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 is determined based on the specific circumstances and projected timelines of the involved Consortium projects.

**Step 3: Recording and Communication of Decision Reasons** Once a decision is 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 the applicant must meet.

# Data retention

Participant data will be retained according to the following guidelines, ensuring compliance with the Protection of Personal Information Act (POPIA) and maximizing its utility for future research:

**Retention Periods:**

* **Original Study Data**: Retained for at least five years after the completion of the project. This includes raw, unprocessed data collected directly from cohort studies and clinical trials.
* **Consortium Shared Data**: Retained indefinitely or for five years post-project or DTA termination, depending on the agreements in place. This data has undergone initial harmonization and is shared among HE²AT Center partners.
* **RP1/RP2 De-Identified Data**: Retained indefinitely. This data has been further processed and de-identified for specific analyses related to RP1 and RP2.
* **Inferential Data**: Retained indefinitely. This includes aggregated and synthetic data derived from the analysis of the other data categories.

**Ongoing Monitoring** After executing a Data Transfer Agreement (DTA), ongoing monitoring ensures compliance with the terms of the agreement, including periodic reviews and audits if necessary.

By adhering to these procedures, the HE²AT Center ensures that data sharing is conducted responsibly, ethically, and in a manner that maximizes the utility of the data for future research while protecting the rights and privacy of data subjects.

# 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 the UCT data analysis platform** | Rodger Duffett | UCT | rodger@csag.uct.ac.za |
| **Managing access to the IBM platform** |  | IBM | sibusisiwe.makhanya@ibm.com |

# 

# Assessment and revision

The Data Management and Analysis Core (DMAC) co-Principal Investigators (co-PIs) will conduct periodic reassessments of the Data Management Plan (DMP) in consultation with the HE²AT Center Steering Committee (SC), including the leads of Research Project 1 (RP1) and Research Project 2 (RP2). These assessments will occur at least every six months to ensure the plan remains effective and up-to-date.

## 16.1. Assessment scope

The reassessment will focus on three key aspects of the data management plan:

1. **Data Process Efficiency**:
   * **Workflow Evaluation**: Assess whether the data processing workflow is functioning effectively and producing data ready for analysis. This includes evaluating each step from data acquisition and harmonisation to storage and indexing.
   * **Error Identification and Resolution**: Identify any issues or bottlenecks in the current workflow and propose solutions to enhance efficiency and data quality.
2. **Compliance and Security**:
   * **Compliance Check**: Ensure that all data management activities comply with relevant legal and ethical standards, including POPIA and the DSI-Africa Data Sharing Guideline guidelines.
   * **Security Measures**: Review and update data security measures to protect personally identifiable information and ensure the integrity and confidentiality of the data.
3. **Usability and Accessibility**:
   * **Data Accessibility**: Evaluate whether the data is easily accessible to authorised users, including partner researchers and members of the HE²AT Center.
   * **Usability for Analysis**: Ensure that the data is in a usable format for analysis, with appropriate metadata and documentation to support effective use by researchers.

## 16.2. Revision process

Based on the findings from the assessment, the DMAC co-PIs will propose revisions to the Data Management Plan. The proposed revisions will undergo the following process:

1. **Proposal Development**:
   * The DMAC co-PIs will draft detailed proposals for necessary updates and changes to the DMP, addressing any identified issues and incorporating feedback from the assessment.
2. **Review and Approval**:
   * The proposed revisions will be presented to the HE²AT Center Steering Committee (SC) for review.
   * The SC, including the RP1 and RP2 leads, will evaluate the proposed changes and provide their approval or request further modifications as needed.
3. **Implementation**:
   * Upon approval, the DMAC team will implement the revised Data Management Plan.
   * All relevant stakeholders will be informed of the changes, and any necessary training or guidance will be provided to ensure smooth implementation.

By conducting regular assessments and making necessary revisions, the HE²AT Center ensures that the Data Management Plan remains robust, effective, and aligned with best practices in data management and analysis.

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:

* Data process: Is the data processing workflow working effectively and resulting in data that is ready for analysis?

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.

# **Annex 1: Key data sources**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name and Source of Dataset | Description | Key Variables | Spatio-Temporal Coverage | Relevance |
| Biomedical Data | | | | |
| Individual Participant Data Platform | Collation of prospectively collected high-quality data from pregnant women & and/or neonates (PROSPERO: CRD42020214637). Data owners: ki platform and NICD repository | Preterm birth, pre-eclampsia, neonatal admission | African cohorts and trials conducted 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) | 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). | 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 | 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) | A global reanalysis dataset combining observed data with the output of meteorological models. | 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 | A global reanalysis dataset combining observed data with the output of meteorological models. Contains hourly data from 1950 to present. | 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) | A global bias-corrected reconstruction of near-surface meteorological variables derived from the ERA5. | Includes a range of surface and near-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 | Temperature and precipitation from different datasets including: GISTEMP, Berkeley Earth, CPC and CPC-CONUS, CHIRPS, IMERG, CMORPH, GPCC and CRU | 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 | 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 | 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 | 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 | 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) | Global elevation data from the Shuttle Radar Topography Mission (SRTM) | Elevation | Released in 2013 | Determination of heat hazard; Urban heat Island Effect |
| High res imagery (ESA Sentinel 2) | Images from the Sentinel 2 satellite pair which view land surface regions in 13 spectral bands. | 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 a requirement to control for pollution effects or to look at combined heat-pollution exposures |
| Aqua MODIS Land Surface Temperature (MYD21A1D & MYD21A1N) | Satellite derived day and night time, high resolution (1KM) land surface temperature dataset. | Land surface temperature | Spatial: Global land surface coverage; 0.00983 deg res. Temporal: daily; 2002/07/04 to present |  |
| Areal/Geospatial Socio-Economic Data | | | | |
| Gauteng City-Region Observatory | GIS raster and shapefiles for the Gauteng City-Region area | 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. |
| General Household Surveys, Statistics South Africa | Annual household Survey | Living circumstances of South African households: education, health, social development, housing, access to services and facilities, food security and agriculture. | Sample survey data, units 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) | Household Survey | 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, units 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 | Distribution of human population (counts and densities) on continuous global raster surface. Input data are extrapolated to produce population estimates for 5-year intervals | 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 (GDELT) | Portion of news coverage about specific area and time related to Covid-19 | 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 HEAT Center project. This includes the Core

[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 unauthorized person.

I further note that I have received appropriate training on my responsibilities and I am subject to professional obligations of confidentiality.

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  ……………………………………. |

# **Annex 4: Data Access Committee: Data Request form**

**Requesting HE2AT Center data**

Proposals for new studies that utilize data collected in the HE2AT Center may be submitted by external investigators in the form of a study proposal at any time. The proposal will be reviewed by the xxx HE2AT Center Committee, using the following criteria:

* **Scientific merit**
  + Study is aligned with HE2AT Center scientific agenda and priorities
  + Research question is scientifically sound and can be tested with the proposed study design and methodologies
  + Appropriate skills and expertise available in the proposed investigators
* **Potential public health impact**
  + Study is relevant to one or more HE2AT Center study populations (infants, pregnant women, postpartum women)
  + Study will answer important public health questions or is in the critical pathway of research toward such answers
* **Feasibility**
  + Study population and variables required are available in the HE2AT Center data
  + Study is feasible within proposed timelines
* **Data management** 
  + The study proposal has outlined how data will be managed to ensure the protection of individual’s health data

Please contact **[xxx](mailto:impaact.capsubmissions@fstrf.org)**with any questions.

All completed submissions should be sent to the HE2AT Center using this email: xxx.

**HE2AT Center Data Request Application Form**

Submit the completed form to the **HE2AT Center** for consideration using this email address: xxx

**GENERAL INFORMATION**

* + 1. **Request submitted by:**
* Name:
* Institution
* Email address:
  + 1. **Date submitted:**
    2. **Title:**
    3. **Lead investigator:**
* Name
* Institution
* Email address:
  + 1. **Co-investigator(s):** Include name(s), institution(s), email(s)
    2. **Relevant studies:** List all HE2AT Center studies from which data are being requested:
    3. **Relevant variables:** List all variables that are being requested:
    4. **Please indicate your preference regarding the type of file format you wish to receive.**

.csv (text)

.xlsx (Excel)

Other (please specify):

* + 1. **For access to HE2AT Center study data, you must have your institution's ethics approval or waiver. Please confirm that you have attached a copy of the ethics approval or waiver from your institution.**

Confirmed

Pending (please provide further details):

* + 1. **For access to HE2AT Center data, we will need to execute a data sharing agreement. Will you adhere to the terms and conditions of the data sharing agreement?**

Yes

No

**STUDY DESIGN**

1. **Background:** Include brief literature review, and any research gaps that the proposed analysis will fill (maximum 200 words)
2. **Study rationale:** (maximum 100 words)
3. **Study aims and objectives:**
4. **Study design and analysis:** (maximum 300 words)
5. **Data management:** Please include information here about your procedures to safely manage and store the data, as well as how the data will be shared between study investigators (maximum 200 words)
6. **Timeline for completion:**
7. **Dissemination plan/impact:** Please indicate meetings, conferences, and/or journals where you are planning to submit this work, with anticipated dates. Please include any other planned dissemination activities (maximum 100 words)

**Annex 5: Terms of Reference for Data Access Committee**

# References

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.

3. Zandbergen, P.A., *Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data.* Adv Med, 2014. **2014**: p. 567049.

1. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [↑](#footnote-ref-1)
2. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [↑](#footnote-ref-2)