## 

HE²AT Center Data Management Plan

## 

## 

Document version number 2.0

Version Date: 19/03/2024

Developed by \_\_Chris 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**](#_heading=h.90unvw71unl0) **2**

[**Scope**](#_heading=h.pc3tnufnw6ch) **4**

[**Purpose**](#_heading=h.i4uz5aavivlp) **4**

[**Definitions and Abbreviations**](#_heading=h.ecv02rfj2as0) **4**

[**Background**](#_heading=h.77uz8igc1q8g) **5**

[**Data categories**](#_heading=h.c4w2mpw9upfe) **7**

[**Data management workflow**](#_heading=h.zesy6oaxx028) **10**

[Setting up the DSA between the health data provider and CSAG/UCT](#_heading=h.2rtxcfwz87xl) 10

[Ethics approval](#_heading=h.5uw6kx5exgt6) 11

[Data encryption and transfer](#_heading=h.roqbrlwemlc0) 11

[Data storage and encryption](#_heading=h.q3h2jluxz5x3) 11

[Data indexing](#_heading=h.cxw4cdht8qlb) 12

[De-identification](#_heading=h.608iy5xsj9cn) 12

[Codebook remapping and harmonization](#_heading=h.75143gdith2w) 12

[Data integration and analysis](#_heading=h.2vbrkk84xn76) 13

[Data analysis platforms](#_heading=h.qs0fwdtkwxpu) 14

[**Data Management, Documentation and Curation**](#_heading=h.r9q5v3g7tpcu) **14**

[Data Security and Confidentiality of Potentially Sensitive Information](#_heading=h.qk2bh718q4fw) **15**

[De-identification](#_heading=h.dd0r6q8foh43) 16

[Data encryption](#_heading=h.rsozytv8qfz6) 17

[Storage isolation](#_heading=h.upsdbat1anjx) 17

[Network firewall and Virtual Private Network](#_heading=h.edojaz9ph8zh) 17

[Local authentication and authorization](#_heading=h.5ap7oa57ymcu) 18

[Data Sharing and Open Access](#_heading=h.ps242ohoj9t6) **18**

[**Procedure for making data available to qualified individuals**](#_heading=h.j1w61tvczmoa) **19**

[**Protection of personal information**](#_heading=h.es8r33s72dlf) **22**

[Data Retention](#_heading=h.kdxc7flfo3p) 23

[Roles and Responsibilities](#_heading=h.2ujnndy4i0vu) **23**

[**Assessment and revision**](#_heading=h.ncatxl7c2igb) **25**

[**Annex 1: Key data sources**](#_heading=h.dl3c6y366sgi) **25**

# 

## 

# 

# Scope

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

# Purpose

The purpose of the DMP is to define procedures and standards for the acquisition, transfer, processing, storage, and access to data within the program. While the DMP includes details of access to data by external parties, data sharing outside of the consortium is further supported by the associated Data Sharing Agreements.

# Definitions and Abbreviations

|  |  |
| --- | --- |
| DMAC | Data Management and Analysis Core of the HE²AT Center |
| SC | HE²AT Center Steering Committee |
| DS-I Africa | NIH Data Science Initiative Africa |
| ELSI | Ethical Legal and Social Implications Projects of DS-I Africa |
| HE²AT Center | Heat and Health in Africa Transdisciplinary Center |
| NIH | US National Institute of Health |
| PI | Principal Investigator |
| Research Hubs | One of the seven NIH DS-I Africa Research hubs |
| TEC | Training and Engagement Core of the HE²AT Center |
| The Program | The HE²AT Center program of activities |
| The Projects | The Research Project 1 and Research project 2 of the HE²AT Center |
| The Cores | The DMAC, TEC and Admin Core of the HE²AT Center |
| Re-analysis | A dynamical model simulation of historical climate evolution continuously nudged by observations in order to provide an approximate historical representation of the climate system |
| Sensitive data | Data that pertains to individuals personal information, health, finances, occupation, etc. |
| Personally identifiable data | Data variables that enable the identification of an individual either directly though names, ID numbers, etc. or indirectly through combining other variables such as locations (GPS, street address), age, gender, and medical information |
| Data Subject | The individuals whose personal information is captured in health datasets |
| Data Provider | The legal entity responsible for authorizing access to a dataset |
| DTA | Data Sharing Agreement |
| DMP | Data Management Plan |
| DAP | Data Analysis Platform |
| LDAP | Lightweight Directory Access Protocol |
| sSA | sub-Saharan Africa |
| WWARN | Worldwide Antimalarial Resistance Network |
| BMGFKi | Bill and Melinda Gates Foundation Ki repository |
| TLS | Transport Layer Security |

# Background

The HEat and HEalth in 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 a population health science approach and applying data science methodologies to improve the health of populations in Africa and beyond. The goal of the HE²AT Center is to advance the development of new health knowledge and human capacities through reusing existing data to generate, and then disseminate heat-health knowledge and innovations.

**RP1 description**

Research project 1 is an Individual participant data (IPD) meta-analysis to assess the size and shape/nature of associations between exposure to high ambient temperatures and selected maternal and child conditions within the first two years of life. Such techniques have not yet been employed in the field of climate change and health, and can overcome many of the limitations of traditional analyses of individual datasets and biases in classic systematic review methodology. The project will systematically identify potentially eligible African cohort studies or trials through systematic mapping reviews. Data will be harmonized through re-coding raw individual participant data into a common set of variables, and subsequently all the individual participants data from each eligible study will be pooled. Analyses which will 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.

**RP2 description**

Rapid urban growth, significant levels of informality and increasingly stretched health services, intersecting with observed past and projected future temperature increases, have resulted in a critical emergent public health challenge in African cities. High ambient temperatures can cause considerable morbidity and mortality in urban areas. The magnitude and pattern of health impacts are determined not just by temperature gradients, but also by geographic, socio-environmental and demographic factors. Understanding this complexity is key to developing effective responses that fit the spatial and demographic heterogeneity of cities. This Project, to take place in Abidjan, Ivory Coast and Johannesburg, South Africa, aims to be the most extensive investigation of the risks posed by heat exposure and urbanity in Africa to date. This will be in aid of developing an early warning system aimed at mitigating the impacts of heat for highly vulnerable groups.

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

In a potential pilot project, social media data will be analyzed to better understand the alignment between people’s perceptions of heat stress and the actual heat exposure drawn from meteorological data. It is envisioned that this analysis will be useful for identifying ambient temperature thresholds that matter for thermal comfort in the study areas. This will inform the development of a target urban heat warning classification system. Various approaches to dissemination information from the early warning system will be considered, including the potential to develop a web-based application. The Project aligns closely with the DS-I Africa objectives, especially its solution focus, possibilities for expansion to other Research Hubs, and potential for progressive expansion to cities across Africa.

# 

# Data categories

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

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

**Health related data is generally considered to be personal data** as it pertains to individuals medical records, diseases history and health events. In some cases health related data is also considered *personally identifiable data* where individuals can be identified either directly through names, ID numbers, etc. or indirectly deduced from address, GPS locations, or cross referencing with other identifiers. Many countries now have specific laws dictating conditions for sharing and processing of such data and associated protective measures to avoid unauthorized access and potential harm to individuals involved in the studies. In South Africa the relevant legistation is the Protection of Personal Information Act (POPIA 2013)[[1]](#footnote-1). This necessitates processing in compliance with the specific requirements of the POPIA act. These requirements and compliance with them is outlined in detail in the section below: POPIA compliance  
  
Due to above mentioned privacy, ethical and legal requirements, health related data acquisition almost always involves an associated DTA between the data holder and the HEAT Center (see data management workflow description below). In particular, where health data is traveling across international borders, compliance with the relevant national laws, permissions, and associated requirements for the sharing and transfer of health data will be followed.

Once data is in South Africa, local ethical clearance and, where data constitutes personal information, POPIA compliance apply.

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

Climate related data will in almost all cases 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 IBM PAIRS data storage and/or CSAG/UCT data storage systems however CSAG/UCT will manage and update the primary data index for climate related data available to the consortium.

All climate datasets used are available through open data policies with no restrictions on non-commercial research use. In some cases there is a requirement to cite the original source.

1. **Remote sensing data**The focus is on data derived from satellite sensors, mainly optical imagery (e.g. satellite images of urban centers) as well as indicators of physical measures such as land surface temperature, soil moisture estimates, vegetation condition, land use and cover, etc.) for the purposes of estimating 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 almost all 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 PAIRS data storage and/or CSAG/UCT data storage systems however CSAG/UCT will manage and update the data index for remote sensing related data

All remote datasets used are available through open data policies with no restrictions on non-commercial research use. In some cases there is a requirement to cite the original source.

1. **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, dwelling type, etc. Typical sources include national census data, and more focused household & demographic survey data.

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

South African census data is already available through the UCT DataFirst data repository and GCRO QoS data is available through the GCRO open data platform as well as through direct queries with GCRO.

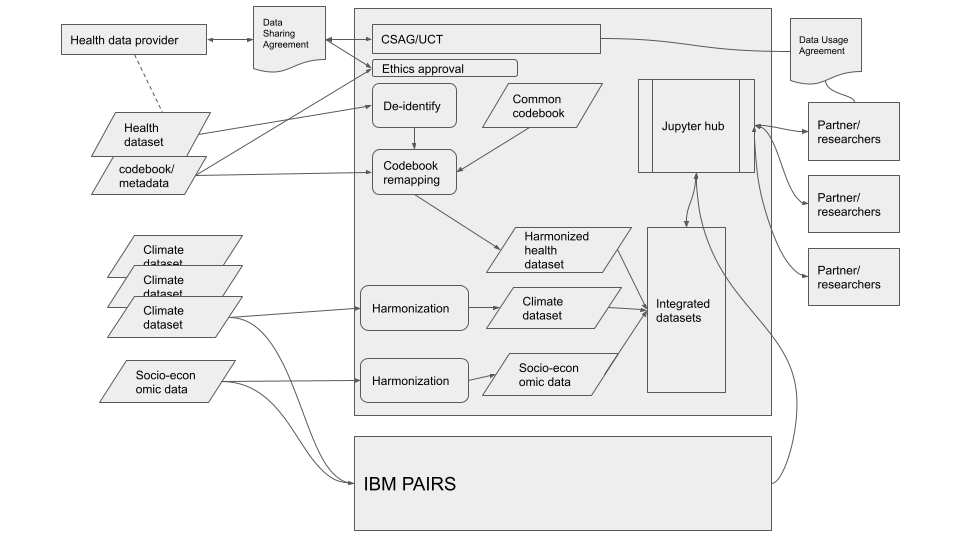
South African census data is aggregated up to small areas and so does not constitute personally identifiable data. GCRO Quality of Life survey data is likewise aggregated to small areas and does not constitute personally identifiable sensitive data.

1. **Social media data**  
   These data represent activity on various social media platforms, initially Twitter. Other data that will be investigated will be the Google Search trends and the Facebook mobility data.  
     
   This is part of a preliminary exploration of the use of social media data to analyze public perceptions. The use of Twitter tweets as data for social media analysis is well described under the Twitter research agreements in order to avoid access or inappropriate use of personally identifiable data.  
     
   Google search trends and mobility data are openly available data and is not associated with individuals. Social media data will not be used in RP1 and RP2 activities, but may be used in a Pilot Project during the course of the study.

# 

# Data management workflow

Figure 1 below maps out the data processing workflow. Details of each element within this workflow are described in detail below.



*Figure 1: Data flow diagram indicating data providers, data sharing agreements and ethics approval, data harmonization steps (aligning with common formats, meta-data etc.), integration, analysis through Jupyter Hub by partners/researchers*

## Setting up the DTA between the health data provider and CSAG/UCT or WRHI

As noted above, a DTA will be required between each health data provider and UCT or WRHI in order to enable the transfer of health data to UCT, and the subsequent processing, analysis, and potentially publishing.

The informed consent process for each dataset that agrees to contribute data to the study will be assessed for ethical considerations. Firstly, we will review informed consent procedures that were undertaken by the participant for the primary study. If a participant has signed a ‘’broad consent’’ for use of their data in future research projects, even though the precise nature of the future research is unclear, this will allow data sharing without further ethical approvals. Participants that have signed ‘’narrow consent’’, where sharing of data is not permitted, and/or “tiered consent”, where the participant chooses to permit data sharing, will be carefully considered. The possibility for reconsenting will be deliberated if no consent for data sharing has been provided by the participant. If reconsenting is not feasible, impossible, or would involve a disproportionate effort, an informed consent waiver will be requested from the Faculty of Science Research Ethics Committee.

There are two main scenarios for DTAs which depends on the requirements of the data provider as well as relevant national or international legislation on health data sharing:  
  
***a) The DTA states that only one person/institution can access/use the data and/or the data must remain on that institutions computing platform/system***  
  
In this case the approach is that the DTA assigns access to the data to UCT and that remote analysis of the data can then be provided through the UCT data analysis platform.  
  
***b) The DTA states that the data is available for use across the consortium or by pre-identified individuals from across the consortium.***   
  
In this case the data should be copied to the UCT DMP as the primary/canonical version, but copies can then be made to be used by consortium partners on their own computing platforms as permitted by the DTA

## Ethics approval

As additional databases become available to the study team and DTAs are agreed to, we will notify the UCT Science Faculty Ethics Committee in writing of the new study that has agreed to contribute data to the study. We will provide information about the database prior to actual data transfer. We will inform the committee of the name of the study, study acronym, contact details of the data owners, details of the ethics approval of the original study including informed consent parameters, study country and provide a copy of the DTA. We will make additional information available on the study where requested by the ethics committee.

## Data encryption and transfer

Once the DTA has been agreed and signed, and no ethical concerns have been raised by the ethics committee, the health data will be transferred to the UCT data platform. If the data constitutes personally identifiable data, it will be encrypted by the data provider and transferred through a data transfer service that utilizes encrypted data transport (TLS). For example, FileSender

## Data storage and encryption

Once transferred to UCT, if the datasets constitutes personally identifiable data, or if stipulated by the DTA, the dataset will be encrypted for storage using 256-bit AES (Advanced Encryption Standard) a standard established by the US NIST (National Institute of Standards and Technology) with encryption keys only available to the minimum number of people required to implement any anonymization or data minimilization process (see team responsibilities below)

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

## Data indexing

Data currently available on the CSAG/UCT data storage system as well as data available on IBM PAIRS will be indexed using an appropriate meta-data standard and this index will be made available by CSAG/UCT data platform (CKAN implementation). The Comprehensive Knowledge Archive Network (CKAN) is an open-source open data platform for publicly sharing and distributing data by public institutions.

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

The current DRS is documented on the CSAG Gitlab Wiki

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

## De-identification

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

## Codebook remapping and harmonization

Codebook remapping involves translating variables in order to align with a common codebook. Each RP and potentially also specific analyses within each RP, will develop common codebooks defining the required health, socio-economic, other environmental, and climate variables. Each source dataset will require a map that defines how variables are translated from source to the harmonized common dataset. For example, pre-term delivery might be translated from the number of days, to a categorical variable. Clinic names might be translated into administrative area identifiers.

Similar to health data harmonization, climate and environmental datasets will be conformed to a common standard through a separate harmonization process. Environmental data (climate data, remote sensing data, socio-economic mapping data) will be homogenized to align with existing meta-data and storage standards (CF conventions and OGC standards) as documented on the Gitlab Wiki

The result of the codebook remapping will enable multiple datasets to be combined into one harmonized dataset including relevant ancillary variables (environmental, climate etc.)

## Data integration and analysis

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

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

1. Basic error screening and correction to identify formatting errors, invalid codes and values, etc.
2. Screening to check for personal identifying information. For example, date of birth is an indirect identifier, implying high risk of re-identification when combined with other indirect identifiers. Date-of-birth can therefore be dropped, after calculating and keeping the age of the patients.
3. Recoding to one of the identified common codebooks. This step will extract variables of interest for each RP, translate variables into the target units including translations of real values into categories, translating from categorisation scheme to another, etc. This step will also include recoding of geographic information. Where point location information is given, the jittering method of perturbing spatial location data will be used, where each point location is off-set in a systematically random way (random distance & random direction). The alternative will be to substitute point locations with grid-cell centroids where the grid-cells correspond to the grid framework to be used for climate & environmental data
4. Further quality control and restructuring of data
5. Integration of climate variables in support of particular analyses

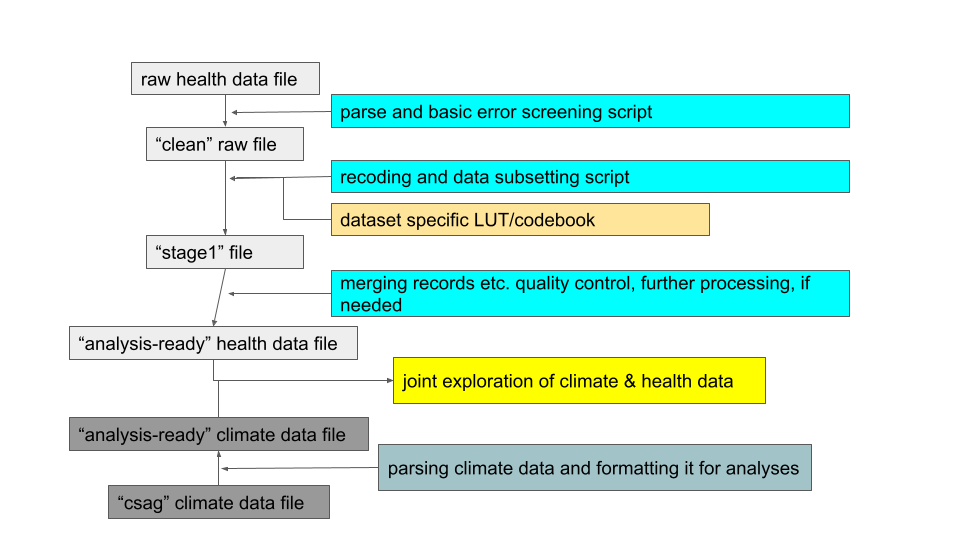


Figure 2: Detailed data processing for health data ingestion

## Data analysis platforms

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

*Juypter Hub* is a collaborative web based Python coding environment that enables analysts to develop analysis code in Python using a web based coding platform. The analysis code is executed on the CSAG HPC platform and datasets stored on the CSAG storage servers are available. The code environment is customized to support all the relevant coding libraries for machine learning (e.g TensorFlow, PyTorch, Keras, Scikit-learn, Numpy, Scipy, Pandas, etc.).

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

# Data Management, Documentation and Curation

Data will be managed by the DMAC team and ***primary data management*** will take place on the CSAG/UCT data platform.

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

Individual partner institutions will have their own data and computing platforms ranging from central institutional platforms through to personal computing devices. Data management within these partner institutions does not fall under this data management plan and we refer to this as ***secondary data management***. However, partners will be encouraged to regularly consider what locally managed data should be integrated into the project wide data management platform, and made available to the broader DSI-Africa programme and beyond. It is critical that health data be kept in a secure and managed platform, therefore where researchers from the various partner institution wish to run analyses on the health data in the CSAG/UCT platform, they will use the DMAC JupyterHub environment from which they will also be able to access data from the PAIRS system through an SDK.

Primary data management will involve:

* Homogenization to agreed data archive standards (see above for each class of data)
* Version control of datasets that are regularly updated to ensure prior version remain accessible
* Harmonizing of health related data
* Meta-data indexing within the CSAG/UCT DMP as well as DS-I Africa ODSP
* Documentation of data on CSAG/UCT GitLab wiki
* Sharing of data management code through CSAG/UCT GitLab code repositories

# POPIA compliance and protection of personal information

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

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

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

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

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

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

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

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

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

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

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

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

## De-identification

POPIA Section 10 prescribes the principle of Minimality which means that *only information relevant to the purpose of processing should be processed*. Where personal information is acquired that is required to fulfill 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[[2]](#footnote-2).

1. Street addresses may be aggregated into geographical regions such that many records map to the same region and it is no longer possible to derive individual residential locations. For example, in RP2, where the highest spatial granularity is needed to map urban heat health outcomes, census small areas or wards with spatial scales of the order of 2 to 5 km will be used to aggregate records. Aggregation will also consider the number of records that map to the same geographical area.
2. Latitude/longitude coordinates may be “jittered” which involves adding random values to each coordinate such that the exact location is lost ( Zandbergen 2014 assesses different approaches to location masking and jittering), while retaining sufficient geographical information to support the analysis. This will likely be the case for RP1 where jittering will be used to shift latitude/longitude locations in the order of 10s of kilometers, adequate to prevent locating individual residential locations.

Geo-location masking/jittering and aggregation will be considered adequate through expert determination involving experts from UCT, IBM and NIH.

## Data encryption

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

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

## Storage isolation

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

## Network firewall and Virtual Private Network

CSAG compute infrastructure sits within the UCT intranet and falls under the broader UCT security policies and services. This ensures that network access to CSAG servers and computer services is carefully controlled and limited to authorized users only.

*UCT broader security policy*

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

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

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

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

## Local authentication and authorization

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

# Data Sharing and Open Access

According to the Research Data Management policy, ‘publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner’. Data is therefore open by default, 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."

Strategies to limit restrictions may include data being de-identified (see above), gaining participant consent for data sharing; and gaining copyright permissions.

**Discoverability**

The HEAT Center and DMAC will implement FAIR principles:

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

# Procedure for making data available to qualified individuals

The SC will develop a Data Access Request Form which people requesting data need to complete prior to consideration of their request. The form will include a proposal outline of the intended research, and the procedures to maintain data confidentiality and security.

The SC will review and approve or reject all requests from the research community, which includes scientists or medical professionals working at academic, non‐profit or government institutions, or commercial companies. We will ensure that all requests conform with NIH policies and procedures including compliance with informed consent procedures if relevant, and any limitations stipulated by the institutions/investigators who contributed data to either of the RPs. The study will comply with the principles of the Data Protection Act of the country of the participating site. Some countries may have restrictions on data sharing outside of the host country, which we will abide by. Researchers who request to share the resources of the HE²AT Center will need to agree to not seek to identify individuals within the dataset, not distribute the data to any other entity, keep the data secure, and acknowledge the HE²AT Center and DS-I Africa as appropriate in publications and presentations (the exact acknowledgement text to be agreed among the DS-I Africa Program). Further, researchers who share the Center’s resources will be strongly encouraged to collaborate with and train African investigators as part of the work they conduct with the resources shared.

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

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

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

The IPD in RP1 involves ‘data sharing’ on a large scale, where we are reliant on the willingness and ability of data owners to share. In that spirit, as the holders of the IPD database it will be beholden on us to share the data with other groups, provided the original data holders agree to sharing and the procedures of HE²AT Center data sharing processes are followed. We will develop formal collaboration agreements around data sharing with the data owners who contribute data to the IPD, which set out the terms and conditions for data reuse. This is important, as for example, some of these investigators may not wish their data to be shared with third parties, for example. While a willingness to share data forms the basis for an IPD and most often reflects a desire of investigators to collaborate, this may not necessarily translate into a willingness to then share data beyond the Center. Sharing of the datasets gathered into the IPD database will thus require signed agreement from the original data owners and may require an application to the local IRB where the original study was done. People who make use of the IPD database will have to undertake to adhere to the authorship guidelines stipulated in the collaboration agreement signed between the HE²AT Center and the research groups who contributed their data.

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

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

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

## Data Retention

Participant data and the data of the children will be retained for a period for at least 5 years after the completion of the project for historical, statistical or research purposes as provided for by POPIA. Appropriate safeguards will be established to secure the records and ensure that the data is not used for any other purpose than the purpose for which the data was originally collected. In any event, no records will be retained for longer than as per applicable regulations.

# Roles and Responsibilities

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Role and responsibilities** | **People** | **Institution** | **Contact** |
| **DMAC PIs**  Responsible for ongoing (quarterly) assessment of data management and changes to the data management plan (annual) | Christopher Jack  Sibusisiwe Makhanya | UCT  IBM | [cjack@csag.uct.ac.za](mailto:cjack@csag.uct.ac.za)  sibusisiwe.makhanya@ibm.com |
| **Health data acquisition**  Identification of relevant health datasets, coordination and development of the DTA | Matthew Chersich for RP2  Craig Parker for RP2  Stanley Luchters for RP1 | WITS RHI  WITS RHI | [MChersich@wrhi.ac.za](mailto:MChersich@wrhi.ac.za)  [cparker@wrhi.ac.za](mailto:cparker@wrhi.ac.za)  stanley.luchters@aku.edu |
| **Data processing and harmonization (including de-identification)**  De-identification, quality control, remapping, harmonization 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  WRHI  WRHI | [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  [Nbrink@wrhi.ac.za](mailto:Nbrink@wrhi.ac.za)  [Cparker@wrhi.ac.za](mailto:Cparker@wrhi.ac.za) |
| **Managing access to UCT data analysis platform** | Rodger Duffett | UCT | rodger@csag.uct.ac.za |
| **Managing access to IBM PAIRS platform** |  | IBM |  |

# 

# Assessment and revision

The Data Management Plan will be periodically assessed by the DMAC co-PIs in consultation with the HEAT Steering Committee including RP1 and RP2 leads. This will take place at least *every 6 months*.

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

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

Based on the assessment, the Data Management Plan will be updated and changes will be implemented. Revisions to the Data Management Plan will be proposed by the DMAC Co-PIs (see above) and approved by the SC.

# Annex 1: Key data sources

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

# 

# Annex 2 : Personal information processing agreement

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

[Full name] hereby agree to comply with the requirements of the POPIA Act of South Africa as regards 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

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

# 

# 

# References

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

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