



The HarmonAlze Toolkit

Automated Health Data Harmonisation and Climate Integration



Natasha Lalloo¹, Craig Parker¹, Rutendo Sibanda², Zororo Chinwadzimba², Perkins Watambwa², Sithembile Musembiri², Tatenda Makanga², Katherine Johnston³, Pierre Kloppers⁴, Lisa van Aardenne⁴, Jemima Effa N'da^{5,6}, Iba Dieudonné Dely^{5,7} & Nicholas Brian Brink^{1,8}, on behalf of the HarmonAlze Consortium

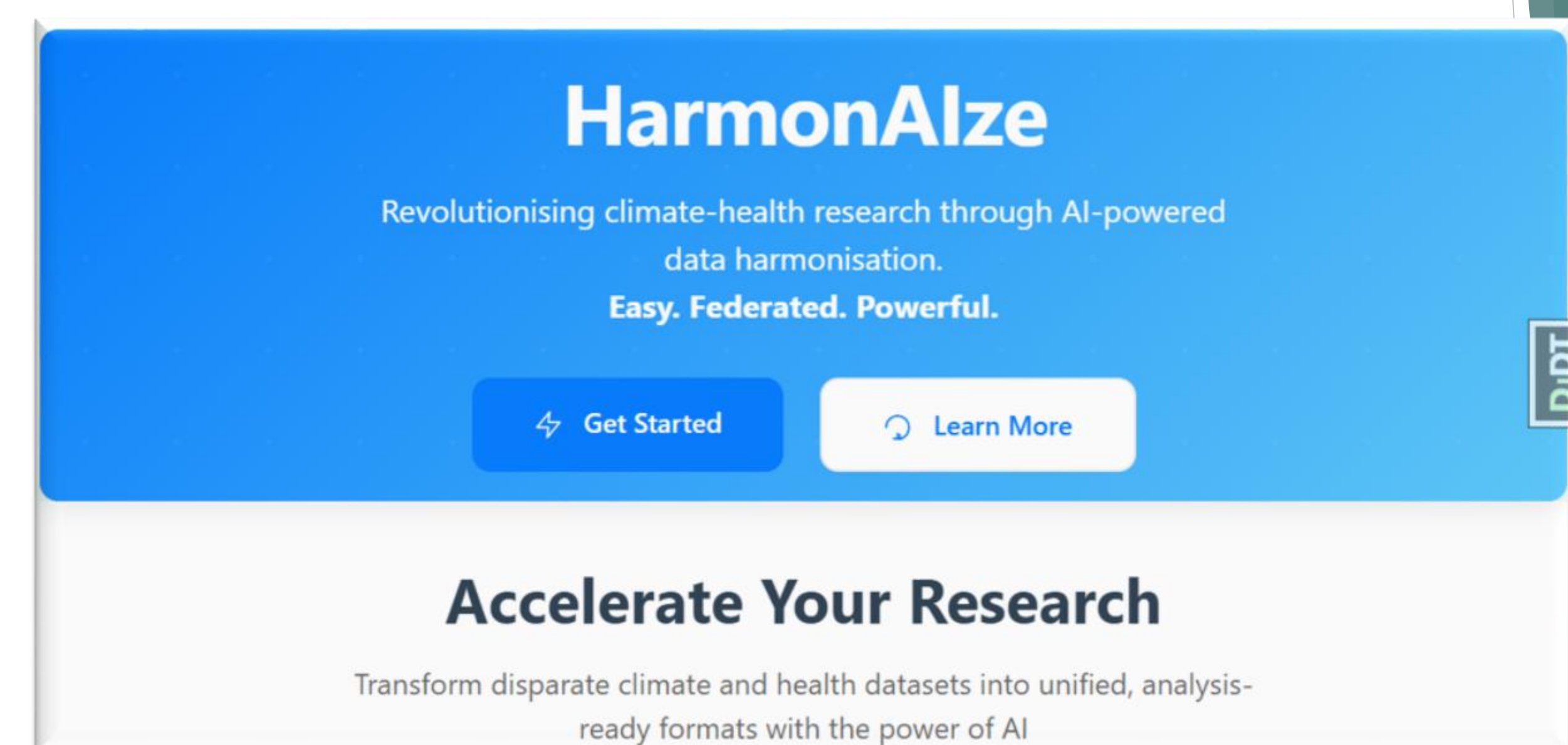
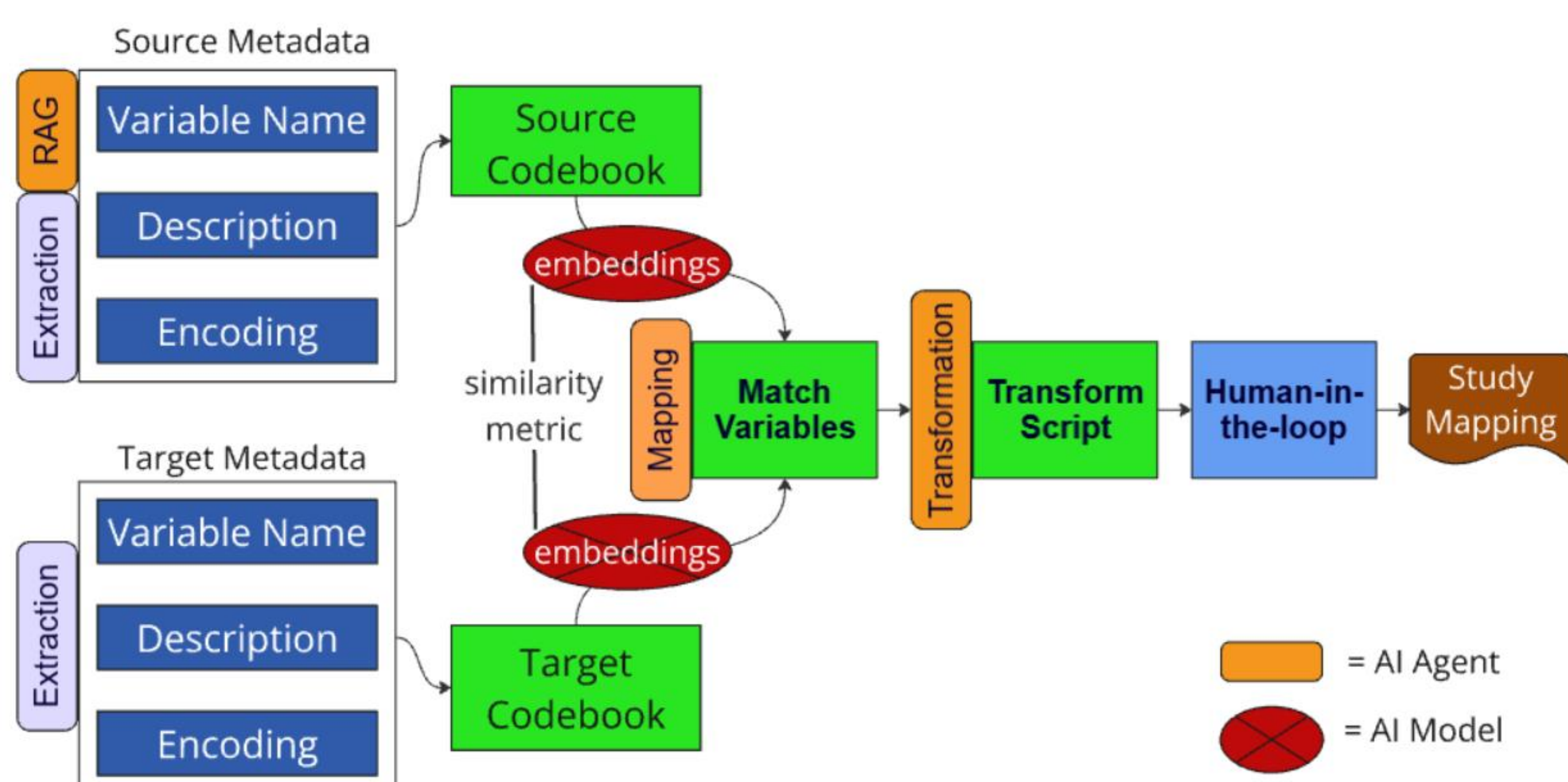
1. Wits Planetary Health Research Division, Faculty of Health Sciences, University of Witwatersrand, Johannesburg, South Africa. **2.** Centre for Sexual Health and HIV/AIDS Research, Harare, Zimbabwe. **3.** eLwazi Open Data Science Platform, University of Cape Town, South Africa. **4.** Climate Systems Analysis Group, University of Cape Town, South Africa. **5.** Centre Suisse de Recherches Scientifiques, Côte d'Ivoire. **6.** Nangui Abrogoua University, Abidjan, Côte d'Ivoire. **7.** University Peleforo Gon Coulibaly, Korhogo, Côte d'Ivoire. **8.** Trinity College Dublin, Dublin, Ireland.

Background

Reanalysis of individual-level health data is a robust evidence synthesis method to generate climate-health insights. However, **data harmonisation** is a **resource-intensive time-consuming undertaking**, and **geolocation** and **climate data linkage** is often challenging.

Solution

A **scalable, semi-automated toolkit** leveraging Artificial Intelligence (AI) to facilitate **data harmonisation**, **geocoding** and **climate integration**



Methods

Codebase: Github for repository, collaboration, version control

Infrastructure: web-based platform built using Django

Privacy preservation: only the data uploader views their own data

Relational database: Entity-Attribute-Value schema

Health data harmonisation: AI-augmented (OpenAI); embeddings-based variable matching and retrieval-augmented generation

Geolocation: semi-automated using Gemini

Climate data linkage: API call through Google Earth Engine, user selects sources (e.g. MODIS, ERA-5)

Results

General

- Ready for **local deployment**
- Early testing positive

Health Module

- User uploads and defines **target codebook** and **study metadata**
- User can **manually verify** uncertain mappings and **input custom transformation code**
- Variables are **successfully harmonised** to target codebook

Geolocation Module

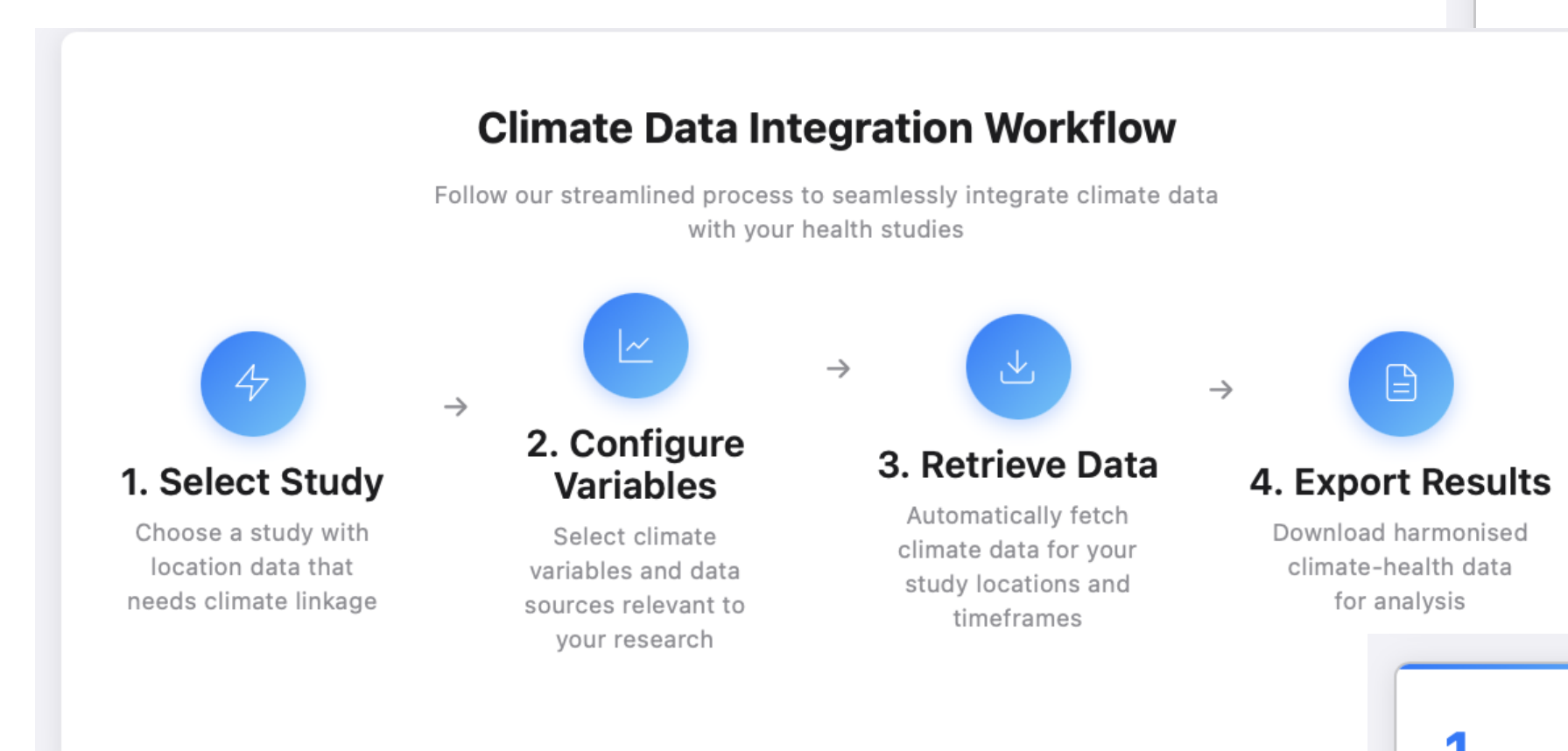
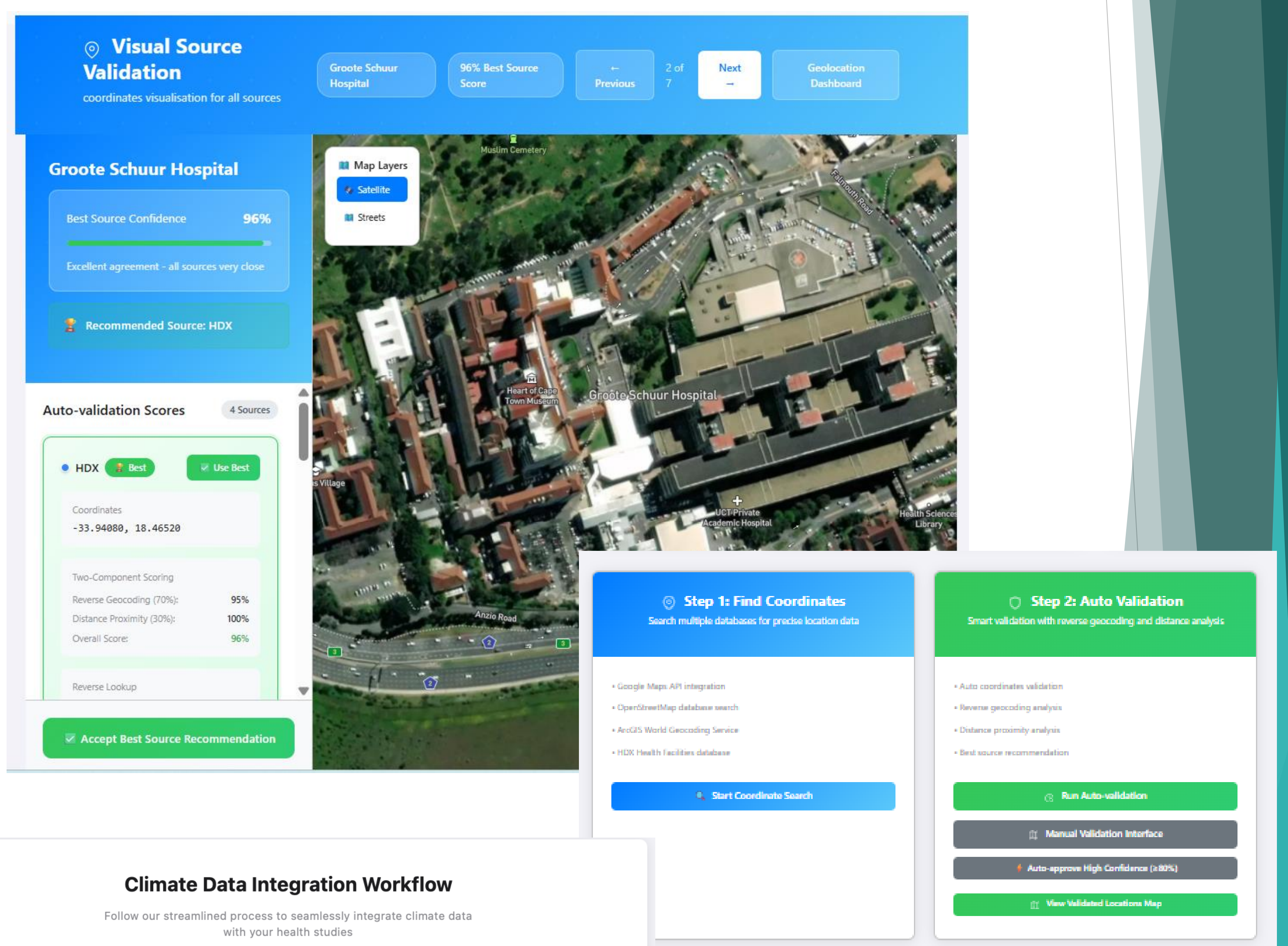
- Locations are **extracted, processed, and validated**
- User can **manually verify uncertain locations** through a user-friendly interface

Climate Module

- User can assign diverse **climate exposure variables**
- Climate variables **retrieved** using standardised date and location

Federated Analysis

- Federated analysis using Datashield preserves sensitive data privacy
- Still a work in progress



Next Steps

- Code review, **validation**, and **cloud deployment**
- Beta testing
- Additional features:
 - Auto-generated codebook
 - Mapping one:many
 - Synthetic data generation
 - "Offline" transformation script
 - Troubleshooting and chatbot

<https://github.com/drnicholasbrink/HarmonAlze>

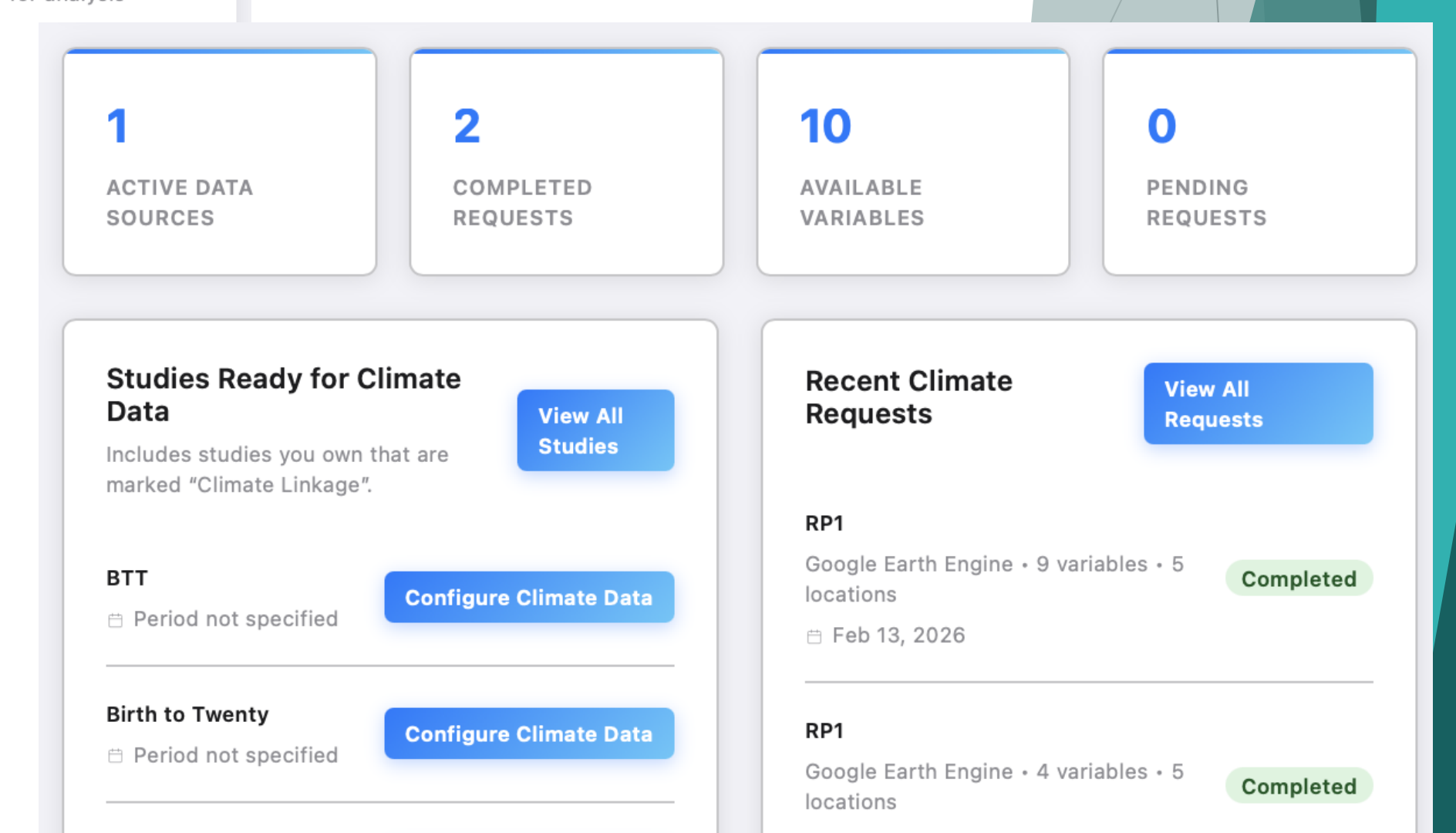


Github Repo



Website

Contact us to collaborate, beta test, or find out more
natasha.lalloo@witsphr.org /
nicholas.brink@witsphr.org
LinkedIn: Natasha Lalloo



Funding: HarmonAlze project is a supplementary grant to the HE2AT Centre, a grant supported by the Fogarty International Center, National Institute of Environmental Health Sciences (NIEHS) and OD/Office of Strategic Coordination (OSC) of the National Institutes of Health under Award Number U54TW012083
Conflict of Interests: Members of HarmonAlze hold investments in the fossil fuel industry through their pension funds. The University of the Witwatersrand holds investments in the fossil fuel industry through their endowments and other financial reserves.
Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

