ROSETTA

Data Management Plan

ROSETTA is co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Project

Fellow Name:	Thompson Gyedu Kwarkye
Project Name	Navigating the Frontiers of Governing Artificial Intelligence in Africa
Project Acronym	AfroGAIN
Project Number	101126578
Supervisor Name	Dr Rónán Kennedy

Data Management Plan

Date	08/07/2025
Version	1.0

1. INTRODUCTION

AfroGAIN investigates how political, sociocultural and ethical processes shape AI governance and what this means for balancing innovation and ethics in Africa. The need for this research is built on AI as a 'double-edged sword' that can help meet Africa's development aspirations while possessing potential risks such as algorithmic biases and human rights concerns. Critics also point to AI's possible negative geopolitical impacts on the continent, particularly as major powers, including the US, EU, and China, compete to establish partnerships within Africa's AI ecosystems, each promoting different standards.[1] While many African countries have been governing AI as a priority, there is limited research on the motivations and politics behind actively pursuing AI governance, including the values, themes, principles, and ethics embedded in the frameworks. AfroGAIN leverages both qualitative and quantitative data from South Africa and Ghana to unpack how AI is building governance processes, how it is envisioned, and the role of local and transnational actors in influencing AI approaches on the continent.

The focus on these two countries is because South Africa dominates Africa's AI governance through its deep learning and data analytics initiatives.[2] This contrasts with Ghana, where no national AI policy exists, even though the country uses AI extensively in health services, agriculture, education, etc.[3] Thus, both countries

[1] Miailhe, N., 2018. The geopolitics of artificial intelligence: The return of empires?. *Politique étrangère*, (3), pp.105-117.

[2] Ade-Ibijola, A. and Okonkwo, C., 2023. Artificial intelligence in Africa: Emerging challenges. In: D.O. Eke, K. Wakunuma and S. Akintoye, eds. *Responsible AI in Africa: Challenges and Opportunities*. Cham: Springer International Publishing, pp.101–117.

[3]Kwarkye, T.G., 2025. *Ghana's Pathway to AI Governance and Its Implications for Africa*. Policy Brief No. 202. Waterloo, Canada: Centre for International Governance Innovation. Available at: https://www.cigionline.org/publications/ghanas-pathway-to-ai-governance-and-its-implications-for-africa/ [Accessed 5 Jun. 2025].

will provide a sample frame to give the research a wider sociocultural dimension across the continent and catalyse information-rich discussions about states at different stages of governing AI. The project is co-funded by LERO, the Research Ireland Centre for Software, the European Commission's Marie Sklodowska Curie Actions COFUND scheme and the University of Galway, with secondment from the African Centre for Economic Transformation (ACET).

This Data Management Plan (DMP) is a living document that will be used throughout AfroGAIN's duration. The aim is to ensure that research outputs remain as open as possible and as closed as necessary, adhering to open science, FAIR principles, and the General Data Protection Regulation (GDPR). This deliverable presents the submitted version (v.1.0) that examines steps in identifying data that will be handled throughout AfroGAIN's lifecycle, including how the data will be created, stored and backed up for future use. This plan also addresses privacy, security and ethical dimensions related to data management. After this version's release, future versions will include refined strategies and additional considerations for the research tools, data structures, and analytical outputs developed by the fellow under AfroGAIN. These revisions may involve input and output data and file format selection changes to reflect the evolving nature of AI governance research and methodological adaptations during the fellowship.

2. DATA SUMMARY

AfroGAIN includes three work packages (WPs), each focusing on a different dimension of AI policy, governance, and stakeholder engagement across Ghana and South Africa. The data collected across these WPs vary in type, format, source, and volume, reflecting the complexity and multi-stakeholder nature of the study as shown in the table below:

Work Package	Type of data	Source	Format	Volume/estimated data size, e.g. bytes, number of objects, files
WP1	Analysed AI policies, strategies, national and international regulations and agile governance tools	Governments of Ghana, Government of South Africa, The African Union, The European Union, Civil Society Organisations, NGOs and other AI Stakeholders in Africa.	.txt .doc .xml	30 documents approx. 10 mb
	Interview transcripts	Unstructured interviews with stakeholders (e.g. ACET, GIZ Fair Forward, and Smart Africa)	.txt .docx .xml	15 interviews approx. 10 mb
WP2	Interview transcript	Semi-structured interviews with state institutions, NGOs, development partners, 'traditional' authorities, diaspora groups, multinational corporations and foreign governments.	.txt .docx .xml	20 interviews approx. 14 mb
WP3	Survey responses	150–300 participants in Ghana and	.csv	150–300 responses,

		South Africa (e.g., policymakers, AI developers, academics and community members).	.xlsx .xml	approx. 10 MB total
	Focus Group transcripts	Up to 10 focus groups with diverse participants (e.g., policymakers, AI developers and community members).	.txt .docx .xml	10 transcripts, approx. 10 MB total

3. FAIR DATA MANAGEMENT

The DMP will ensure FAIR principles (Findable, Accessible, Interoperable, Reusable) by (i) publishing the data in Zenodo and the project's Observatory (www.afrogain.com), using persistent identifiers and structuring metadata to ensure data are reliably found: (ii) uploading transcripts, field notes, and analysed scripts to GitHub to allow access after applying anonymisation techniques to remove personally identifiable information and other sensitive elements: (iii) saving the transcripts in XML to ensure interoperability with other data systems: (iv) documenting the steps for data collection, including sampling approaches and participant selection, on the Observatory to facilitate long-term reuse. The fellow will also include a codebook and a methodology section in the dataset documentation. These are further explained below.

a. Making data findable, including provisions for metadata

Deposit the data in a repository: The complete dataset—including data from the content analysis, interview and focus group discussion (FGD) transcripts, and survey responses—will be made available on Zenodo. Zenodo assigns metadata to each dataset, making it easier for researchers to find data through search engines and academic databases. Additionally, all deliverables will be stored on AfroGAIN's AI Governance Observatory at www.afrogain.com. Both Zenodo and the observatory will maintain an indexed and searchable catalogue of all shared data to enable users to filter datasets by variables such as country, data type, research theme, or collection method.

Use persistent identifiers: To enhance the discoverability and accessibility of research outputs across academic platforms, all datasets deposited in Zenodo and the project's observatory will be assigned persistent identifiers such as Digital Object Identifiers (DOIs). These identifiers will ensure stable, long-term links to the data, enabling accurate citation and reliable retrieval.

Metadata Structuring: Each dataset's metadata will be structured using standard schemas (e.g., Dublin Core or DataCite), allowing seamless integration with academic search engines and AfroGAIN's AI Governance Observatory. This approach not only supports transparency and reproducibility but also increases the visibility and impact of the research.

b. Making data accessible

Upload to GitHub: Apart from depositing the data in Zenodo and the observatory (which also promotes accessibility), transcripts, field notes, and analysed scripts will be uploaded to GitHub. Before upload, all materials will undergo a rigorous anonymisation process to remove personally identifiable information and other sensitive content that could compromise participant confidentiality. This is particularly important given

that some data may involve vulnerable or politically exposed groups. The anonymisation process will follow GDPR- compliant protocols and will be reviewed in collaboration with the project supervisors and the Research Data Steward at the University of Galway. Once anonymised, the datasets will be version-controlled and shared under appropriate open licenses to facilitate accessibility.

Accompanying README, Detailed Codebook and Metadata: Each dataset will be accompanied by comprehensive documentation, including a README file, a detailed codebook, and structured metadata. The README file will provide an overview of the dataset's purpose, contents, and usage instructions. The codebook will explain variables, coding schemes, and any transformations applied to the data, ensuring transparency and interpretability. The metadata will use standardised schemas to describe the dataset's structure, content, and contextual background, supporting effective data discovery, reuse, and integration into broader research frameworks.

Publish in open-access journals: AfroGAIN will engage in open peer review by publishing in peer-reviewed journals with open access and using Open Research Europe, where manuscripts will be publicly reviewed and responded to. By embracing open peer review, the fellow aims to contribute to a more inclusive and dynamic exchange of knowledge, enhancing the impact and reach of AfroGAIN.

c. Making data interoperable

Saving Files in Extensible Markup Language (XML): All transcripts will be saved in XML format to ensure interoperability with a range of data systems. XML is widely recognised for its ability to structure data in a consistent and machine-readable way, making it ideal for integrating with digital repositories, qualitative analysis software, and institutional databases.[4] Using XML aligns with best practices in open data management and ensures that the data is interoperable for researchers across disciplines. Saving transcripts in XML further allows for the direct addition of metadata within the file structure. This metadata will document crucial information such as interview dates, locations, participant roles, anonymisation status, and coding references. This format facilitates archiving, retrieving, and linking the qualitative data to associated materials such as survey datasets or field notes, reinforcing the project's commitment to interoperability.

d. Increase data re-use

Documenting each data collection process: AfroGAIN will document each step in the project's observatory, Zenodo and GitHub. This documentation will cover the full methodological lifecycle, including sampling strategies, recruitment procedures, inclusion and exclusion criteria, and ethical considerations. By openly sharing this information, the project will enable future researchers, policymakers, and institutions to understand the context and rigour behind the data, assess its relevance for secondary analysis, and replicate or adapt the approach in comparable settings.

[4] Coyle, F.P., 2002. *XML, Web services, and the data revolution*. Addison-Wesley Professional.

Detailed Codebook and Methodology Section: In addition, each dataset uploaded will include a detailed codebook and a methodology section. The codebook will define variables, categories, and codes used during content and thematic analysis, while the methodology section will detail research objectives, instruments, and analytical techniques employed. These components will be critical for users to interpret the data accurately and apply it responsibly.

4. OTHER RESEARCH OUTPUTS

In addition to datasets, AfroGAIN will generate several digital research outputs, including coding schemes, analytical workflows (e.g. NVivo and SPSS project files), survey instruments, interview protocols, and visualisations of governance frameworks. These outputs will be version-controlled and published on the project's observatory, Zenodo and GitHub under permissive open-source licenses (e.g. MIT or CC-BY) where appropriate. Each file will be accompanied by metadata following the DataCite schema. Workflow documentation and protocols will also be integrated into the AfroGAIN's AI Observatory to facilitate replication and secondary analysis by researchers and practitioners.

Where physical outputs arise, such as printed survey forms collected in low-digital-access areas, they will be digitised, anonymised, and securely destroyed after transcription and storage following ethical guidelines. Any physical notes or consent forms will be securely stored during fieldwork and transported only under encryption-compliant conditions, with scanned digital versions retained for institutional archiving. This will ensure that all research materials, whether digital or physical, are documented, ethically managed, and, where possible, made openly available.

5. ALLOCATION OF RESOURCES

a. Costs for Making Data FAIR

The estimated costs for making data FAIR were considered during the application phase of this project. These expected costs include, but are not limited to, the following: (i) secure data storage during and after the project lifecycle (e.g. cloud servers, encrypted external drives, hosting and maintaining the Observatory); (ii) data curation and anonymisation services (especially for sensitive interview and FGD data); (iii) publishing and archiving fees; (iv) version control and maintenance of the GitHub and Zenodo repositories; (v) translation services for multilingual survey data (English, Akan, Xhosa) and (vi) metadata preparation and documentation efforts (e.g. codebooks and XML structuring). These direct and indirect costs are estimated at €1800 across the 24-month project timeline.

b. Cost Coverage

All data management and Open Science-related expenses are covered under the €4920 research budget allocated to the fellow as part of ROSETTA. This budget supports a range of eligible costs, including repository fees, secure data storage using cloud infrastructure or encrypted drives, data anonymisation and curation services, and metadata preparation such as codebooks and XML structuring. It also covers software and tools for version control (e.g., GitHub, NVIVO, SPSS), translation and documentation for multilingual outputs, and expenses related to maintaining the online observatory.

c. Responsibility for Data Management

The primary responsibility for data management lies with the fellow, who will coordinate all data collection, processing, documentation, and sharing activities. The fellow will work with the project supervisors and the Research Data Steward at the University of Galway to ensure GDPR compliance and ethical integrity. This DMP should be ready in the second month of the project's timeline (July 2025) and updated as necessary.

d. Long-term Preservation Strategy

Long-term preservation of AfroGAIN's data will be ensured through Zenodo and GitHub. These platforms guarantee stable access and archiving for at least 10 years, with persistent identifiers (DOIs) ensuring discoverability. A selection process will determine which data has long-term reuse based on impact, ethical clearance, and potential for policy or academic application. Decisions on data retention will involve the fellow, supervisors and the Research Data Steward, ensuring only anonymised and ethically shareable content is preserved. Sensitive raw data (e.g. original recordings) will be securely deleted after transcription and quality assurance to reduce privacy risks.

6. DATA SECURITY

Given the sensitive nature of the information collected, from interview transcripts to focus group discussions and policy documents, robust measures have been implemented to ensure data integrity, confidentiality, and compliance with ethical and legal standards, including the GDPR. As described above, each dataset is stored using secure, institutionally supported infrastructure, backed up regularly, and protected through encryption, anonymisation, and restricted access. The table below summarises these procedures across the three work packages.

Work Package	Type of data	Storage	Backup	Security and Protection	Transfer of sensitive data
WP1	Analysed AI policies, strategies, national and internationa l regulations and agile governance tools and interview transcripts	Approved cloud servers such as Zenodo and GitHub after anonymisation); local encrypted storage during fieldwork	Regular automatic backups on institutional servers and external encrypted drives	GDPR-compliant anonymisation; pseudonymisation; ethical approval required; secure login access and encrypted data storage	Encrypted transfers via secure cloud platforms or institutional VPN; direct handover avoided unless encrypted and necessary
WP2	Semi-structured interview transcripts	Encrypted local devices during fieldwork; University of Galway approved servers; Project Observatory, GitHub (post-	Daily backups to encrypted external drives during fieldwork; regular institutional	Interviews anonymised and stored using coded identifiers; access restricted to fellow and supervisor; compliance with GDPR and national	Secure encrypted uploads; institutional VPN or password-protected cloud sharing only

		anonymisation)	cloud backups	ethics protocols	
WP3	Survey responses; Focus Group Discussion transcripts	Institutional SPSS/NVivo database; GitHub for anonymised versions; secure university file storage	Redundant institutional backups; encrypted storage for raw data and regular GitHub and Zenodo updates	Multilingual data pseudonymised; survey responses coded to prevent re-identification; data handling supervised by the Research Data Steward	Surveys conducted on encrypted platforms (e.g., SurveyMonkey); FGDs transferred via secure audio upload and transcribed locally under encryption

7. ETHICS

a. Informed Consent and Data Sharing:

AfroGAIN ensures that informed consent for data sharing and long-term data preservation is fully integrated into all research instruments involving personal data collection. Participants, policymakers, AI practitioners, civil society actors, and the general public will be informed about the nature and purpose of the research, how their data will be used, who will have access, and the duration of data retention. Consent forms will include data sharing provisions, and participants will be informed of their right to withdraw at any time during the data collection phase without consequence. Consent materials and questionnaires will be made available in English, Twi (Ghana), and Xhosa (South Africa) to address literacy and language challenges, accompanied by oral explanations where necessary.

b. Anonymisation and Sensitive Data Protection

To ensure data diversity, participants will be recruited through extensive collaborations with secondment host and local stakeholders to develop a sampling frame that allows for the equal selection of women, men, LGBTQIA+ groups, ethnic minorities, etc. However, the fellow is aware that some parts of the research areas are not open to LGBTQIA+ groups, with some facing potential persecution if identified. Given the sociopolitical sensitivities and the potential for harm to participants, AfroGAIN will adopt strict anonymisation and pseudonymisation protocols. Personally identifiable information will be removed or replaced with pseudonyms. The results will be presented in aggregated formats to prevent re-identification. This is particularly important in dealing with politically sensitive topics and marginalised groups.

Collected data will be stored securely using encrypted digital systems (e.g., Secure File Transfer Protocol) for transmission and locked physical storage for hard copies, with access restricted solely to authorised personnel. These measures will regularly be reviewed with the project supervisors and the ethics oversight institutions at the University of Galway to ensure adherence to evolving standards.

c. Legal and Ethical Compliance

AfroGAIN will comply with the GDPR and local data protection laws in Ghana and South Africa. Ethical

oversight will be a continuous process, beginning with the fellow's familiarisation with the University of Galway's ethical approval process in M1 and M2 (June- July 2025). Formal ethics approval will be sought before fieldwork or human data collection in M3 (August 2025). In parallel, applications for national ethics approvals are underway with the University of Cape Town and Ghana's Centre for Scientific and Industrial Research. All developments related to ethics clearance will be documented within the project's progress reports and revised versions of the DMP.

d. Inclusion and Risk Mitigation

AfroGAIN will take deliberate steps to ensure the safe and ethical inclusion of underrepresented and at-risk populations, such as LGBTQIA+ individuals, non-binary people, ethnic minorities, and people with disabilities. Established local networks will support recruitment to ensure culturally appropriate outreach and sampling. In contexts where disclosure may expose participants to risk, the fellow will employ careful sampling strategies, data grouping, and anonymisation techniques to preserve confidentiality. Prior experience and existing relationships in the research sites will enhance the fellow's ability to navigate these risks responsibly, with additional guidance drawn from ethical scholarship on inclusion and safety.

Researching in challenging regions with complex sociopolitical issues presents a significant security risk and the potential for delays in data collection. The fellow is well-prepared, having previously conducted research in these regions. He will leverage existing networks with institutions like the Universities of Cape Town and Ghana and organisations such as Irish Aid, ACET, and GIZ Fair Forward to remain informed of on-ground realities and mitigate data access issues. To further reduce these risks, flexible timelines will be adopted, and digital platforms like Zoom and SurveyMonkey will be employed to continue data collection if travel becomes restricted. As a contingency, if in-person data collection becomes impossible, the fellow will pivot to countries with similar sociopolitical dynamics, for example, substituting Ghana with the Ivory Coast.

Delays in ethical approvals could impede research progression. To pre-empt this, the fellow will engage with ethical review boards and relevant authorities in months 1 and 2 to understand their requirements and approval timelines. Ethical applications will be submitted to multiple review boards, including those at local universities to enhance the probability of timely clearance. Should delays still occur, the fellow will pivot to other productive tasks, such as drafting the non-empirical components of academic papers.

Difficulty in securing a sufficient number of participants could limit data reliability. This will be managed by leveraging existing relationships with local organisations and universities from the onset. A multi-pronged recruitment strategy will be employed, involving social media, local events, and networks of both the fellow and secondment hosts. Virtual meetings will address scheduling challenges. If recruitment proves inadequate, the inclusion criteria may be broadened or the geographic scope extended as a contingency measure.

8. OTHER ISSUES

Benefit Sharing and Capacity Building

AfroGAIN will embed capacity-building and benefit-sharing actions to promote equitable outcomes in its

engagement with Ghana as a lower-middle-income country. This includes AI governance and ethics training initiatives for local researchers, civil society, and policymakers. By partnering with ACET as the secondment host for AfroGAIN, the project will facilitate knowledge exchange and support the development of locally relevant governance frameworks. A dissemination, exploitation, and communication plan will reinforce these activities, including public workshops, policy briefs, and contributions to an AI Governance Observatory website. This will ensure that the project's findings serve the public interest in developing economies and contribute to long-term institutional strengthening on the continent.

9. HISTORY OF CHANGES

Version	Date	Change	Author	Role
0.1	05/06/2025	Initial draft	Thompson Kwarkye	Research Fellow
1.0	01/07/2025	Submitted version	Thompson Kwarkye	Research Fellow