**Interim Progress Review for Junior Researchers**

**Programme: Africa AI Research Award (AAIRA)**

Project Title: ***Unlocking Biomedical Data for AI Health Research in Africa Using GeneNetwork as an Example***

Name of Junior Researcher: ***Bonface Munyoki Kilyungi***

Name of Supervisor: ***Dr. Pjotr Prins, Dr. Shelby Solomon Darnell, Dr. George Githinji***

Institution: ***Strathmore University***

1. **Key achievements**

Briefly present what you think are the 2 or 3 biggest achievements or results that happened this last year. Explain why you think these are the most important results or achievements and their contributions to the field of AI4D.

|  |
| --- |
| - ***Self-Documenting Domain Specific Language from SQL to RDF:*** Creating a self-documenting DSL that parses SQL to RDF declaratively addresses a key challenge in AI4D which often involves dealing with complex and diverse data sources, usually stored in SQL databases. This DSL simplifies the process of converting SQL into a more semantically rich format---RDF---as well as allowing for the addition of extra ontologies while doing so. This simplification can be useful for development projects in resource constrained environments where expertise in data manipulation may be limited. By making this process declarative and self-documenting, and limiting learning to only a small syntax, it allows less experienced people to be able to transform data in a meaningful way.  - ***Modeling Omics Data from Genenetwork with RDF:*** The use of the aforementioned DSL to model omics data from Genenetwork and hosting it on a public virtuoso instance is a significant step towards advancing AI in healthcare and life sciences. Genomic and omics data are critical for disease research and precision medicine. Representing this data as RDF allows it to be more accessible to machines. Publicly hosting this data on Virtuoso ensures that promotes collaboration, knowledge sharing and innovation in the health sciences fields. |

1. **Environmental sustainability**

Describe how your project integrates contributions for environmental sustainability in the implementation of its activities. Discuss strategies you put in place that contribute to reduce your carbon’s footprint and ensure an efficient use of resources and energy. For projects of the innovation pillar, discuss their potential to tackle negative impacts of climate change and environmental degradation. What are some emerging learnings?

|  |
| --- |
| N/A---This project mostly entailed creating a Domain Specific Language for parsing data. |

1. **Carbon footprint reporting for model development**

If your project (or sub-projects) has trained a model, please report on the carbon footprint of that training process. We suggest using: <https://codecarbon.io/> but other tools can be used.

|  |
| --- |
| N/A---No models were trained. |

1. **Innovations to address Africa’s development challenges**

Describe your home-grown AI innovation developed, tested or scaled through your project. Describe the current stage and how they can contribute to inclusive and responsible AI practices and can benefit marginalized populations.

|  |
| --- |
| This project involved creating a Domain-Specific Language tailored for use by researchers in Africa. Its primary purpose is to facilitate the transformation of their SQL data into RDF format while allowing for additional annotations as needed. Currently, the DSL is feature-complete. However, there is room for further enhancements, particularly in accommodating various file formats such as JSON, CSV, and different database types. Additionally, apart from the DSL for data transformation, there is a desire to create a declarative language for querying RDF data, with a similar approach to the one presented [here](https://www.swi-prolog.org/pldoc/doc_for?object=section('packages/semweb.html')). |

How has this project contributed to a deepened understanding of the social impacts (positive and negative) of AI innovations? Has it generated lessons learned and best practices in how to design and scale responsible AI4D innovations in the African context? How are these applied or used?

|  |
| --- |
| This project has attempted to address the challanges of sharing data in a FAIR way. By providing and storing data in RDF as a semantic graph, it has made data sharing more accessible and meaniningful. Furthermore, this project has allowed for data annotation with ontologies thus enhancing research quality and promoting the use of FAIR data. Using and enforcing FAIR data will facilitate the exchange of data and insights among different organisations, thus fosterning responsible ML/AI deevelopment. |

Provide the number of private sector actors (start-ups, consultancies, private companies) involved in the development, testing or scaling up of innovations. Specify their name and how they are involved.

|  |
| --- |
| - The University of Tennessee Health Science Center --- This is the organisation that provided GeneNetwork data for use during the creating and provision data.  - KEMRI Welcome Trust --- This is the organisation that will help in hosting a server to host RDF data for use here in Kenya. |

1. **Policy and regulations that promote a responsible and inclusive AI**

Has this project informed and facilitated the development or the revision of public policies, legislations and regulations that promote the inclusive benefits of AI? It is not necessary to demonstrate direct attribution to your project; if relevant focus on your project’s contribution to policy or regulations that support responsible AI.

|  |
| --- |
| No |

Does this project have stories of policy and decision makers actively engaged in discussing AI4D issues at policy or other events? If yes, please provide names, dates and contexts:

|  |
| --- |
| No |

1. **Amplifying African AI talent**

Describe and indicate the number of learning activities that your organisation put in place to strengthen capacities of staff, partners and beneficiaries as part of this project. These include internal and external activities, such as workshops, lectures, trainings, coaching services, etc. Where possible, specify the number of participants and the female-male ratio.

|  |
| --- |
| - One planeterium hosted on 30TH January 2023. All people who attended were male.  - [GeneNetwork Summer Of Code](https://issues.genenetwork.org/topics/biohackathon/GNGSoC2023): This project had a footprint on all the sub-teams involved. There were (active participants): 8 Male and 2 Females. |

Provide the number of male and female graduate students (Masters and PhD students) who have participated in online and offline training and support. Specify the female-male ratio, noting the objective is at least 50% women.

|  |
| --- |
| N/A. No online nor offline training has been done. |

Has this project helped team members or affiliated researchers to be recognized as thought-leaders and called upon for their inputs? If yes, please provide names, dates and contexts.

|  |
| --- |
| N/A. |

Have you contributed to the development of an inclusive AI workforce (practitioners, private sector actors, decision makers) on the continent across multiple sectors and with diverse skills in the loop of AI systems? Please list examples of machine learning engineers, policy makers, private sector and civil society actors that are contributing to the development of a vibrant responsible AI4D community.

|  |
| --- |
| This project has resulted in the recruitment of two people from Kenya to the GeneNetwork open source project: *Priscilla Muli* and *Alex Mwangi*. Priscilla is actively involved in the development of an internal Large Language Model (LLM) designed for querying Genenetwork, allowing researchers to engage in conversational interactions akin to ChatGPT. A notable aspect of this LLM model is its utilization of RDF metadata as input. On the other hand, Alex is primarily focused on the integration of RDF metadata into the broader Genenetwork ecosystem, in particular doing some Machine Learning tasks such as improving correlations on exceptionally large datasets.  Also, this project has motivated [Fahamu AI](https://fahamuai.com/), an african start-up lead by Adrian Kibet and Brian Muhia, to work on upgrading their internal LLM to be able to ingest RDF metadata from labs and other universities in order to broaden their market reach. |

1. **Publications and papers**

List research outputs that are publicly available as direct results of research supported by your project. Examples include peer-reviewed publications, conference papers, datasets, working papers, articles, book chapters and reports.

*There are no papers at the moment. However, the following titles are under preparation:*

*- "GPT3 In the pursuit to easily query Genomic Knowledge" by S. Solomon Darnell, UTHSC Department of Genetics, Genomics and Informatics (GGI)*

*- "Developing a Domain-Specific Language for AI Applications in Biological Data Analysis" by Bonface Munyoki K., Strathmore University*

*Also, my Master’s* [*dissertation*](https://github.com/BonfaceKilz/dissertation) *will be based off this project.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Title** | **Author (s) and institution(s)** | **Author (s)’ gender** | **Venue (journal, book, series, etc.)** | **Date submitted/ accepted/ published** | **Open access?** | **Peer-review?** | **Uploaded to IDRC Connect**  **?** | **Link (if available online)** |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

**8. Outreach tools and processes**

List knowledge outputs that are focused on knowledge translation and public awareness. These include stories developed, policy and knowledge briefs, brochures, blogs, op-eds and other non-academic products.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type (policy brief, blog, op-eds, brochure, etc.)** | **Name of publication** | **Date** | **Reach (global, regional, national, local)** | **Focus of tools** | **Link (if available online)** |
| Repository | N/A | N/A | Global | The main DSL responsible for transforming data | https://git.genenetwork.org/gn-transform-databases/ |
| Repository | N/A | N/A | Global | Versioned auto-generated documentation for the DSL’s output | https://github.com/genenetwork/gn-docs/tree/master/rdf-documentation |
| Website | N/A | N/A | Global | A public website for querying Genenetwork metadata | https://sparql.genenetwork.org/sparql |

**9. Media coverage and citations in public events**   
Please provide details of any media coverage (radio, television, print media, etc.) of your research activities or outputs. This can also include references which cite or otherwise reference your work, including speeches, public statements, policy documents and regional and global forums where your work is featured or discussed.

**N/A**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type (radio, speech, interview, etc.)** | **Name of publication, network, etc.** | **Date** | **Reach (global, regional, national, local)** | **Focus of article or broadcast** | **Link (if available online)** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |