Unveiling Synergies: Mapping the Nexus between NIHR BioResource Contributions and Scientific Publications

Olawale Ogundeji[1]

[1]NIHR BioResource for Translational Research; Correspondence: waleabiodun85@gmail.com

Background of the Study

- In biomedical research, data resources and resulting scientific publications frequently require collaborative efforts and public engagement
  - The BioResource is a key organisation in this ecosystem, consistently authoring and acknowledging publications that result from its volunteers' participation
- This project aims to enrich and visualise BioResource publications data, to help characterise and convey research outcomes, and support existing public engagement
- As the project progresses, linking publications to their respective studies and data access activities will enhance transparency, allowing participants to better understand the impact of their contributions
- Using data from a corpus of 400+ publications, along with insights from the operations team and data science tools/AI technologies, we propose to create
  - a comprehensive database capturing the origins of each publication; along with
  - visualisation and textual output highlighting the BioResource's impact
- This poster showcases one of those aspects of the project: the potential for innovative visualisations highlighting the collaborations and advancements facilitated by the BioResource

Evolution

In the project's current phase, we have:

- Enriched existing publications data with full text, and cleaned existing data
- Developed visualisations to better characterise and understand the corpus
- Used various AI techniques to explore the full texts, including analysis of topics, and interconnections among scientific contributors

In the next phase, the work aspires to:

- Create a database that highlights the origins of each publication
- Further explore the generation of lay or public benefit summaries for external and internal audiences
- Develop a visual narrative of the BioResource's positive impact

Progress So Far

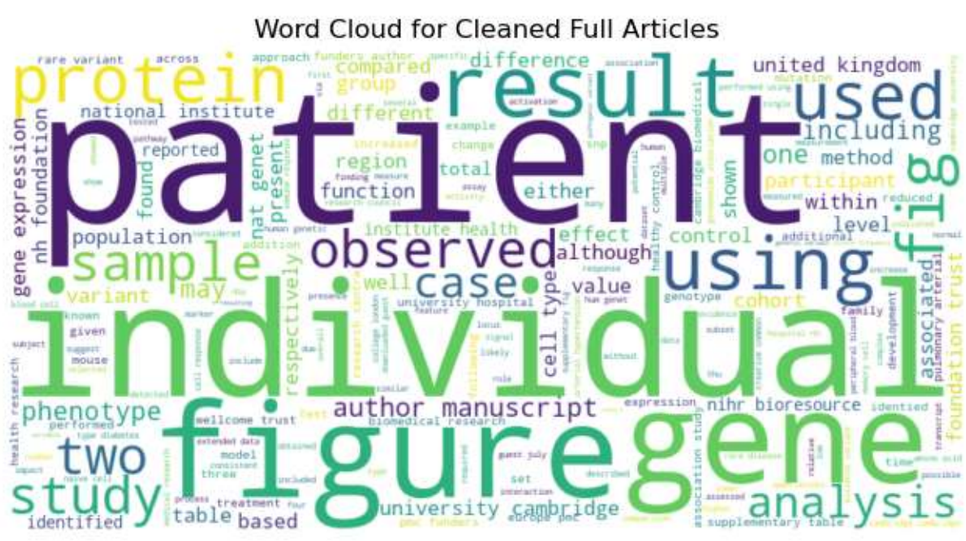

Fig 1.0: An example word cloud of publications using full text

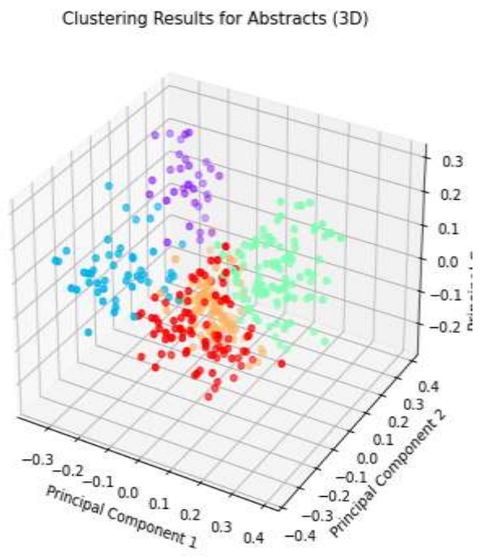

Fig 2.0: Visualisation of clustering output using publication abstracts

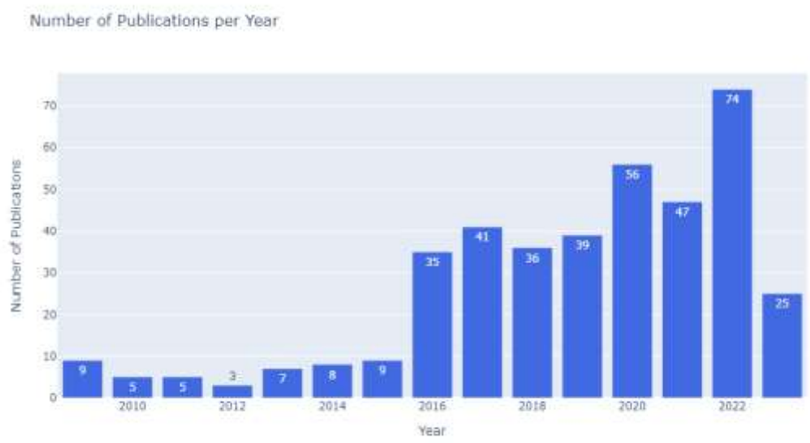

Fig 3.0: Further visualising the publications over time

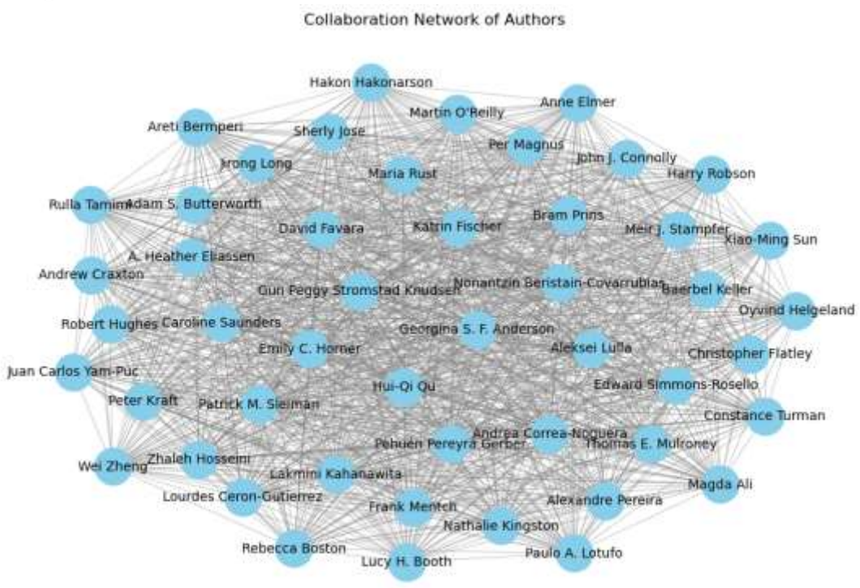

Fig 4.0: Illustrative graph of network analysis using a subset of authors

Research Methods

The project has employed a variety of methodologies in characterising and visualising the dataset so far:

1. **Visualisation:** Word clouds and tools like Plotly help initial understanding of prominent themes and keywords
2. **Clustering Analysis:** We explore groupings of the publications using k-means clustering, towards understanding shared themes or content, and the diversity and distribution of research topics
3. **Topical Modelling with NMF and LDA:** Through Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA), we extract latent topics deepening understanding of research themes
4. **Network Analysis:** We use networks graphs to highlight the interconnectedness in publications through their authors
5. **Natural Language Processing (NLP) Techniques:** Leveraging text analysis, language detection, and named entity recognition, we identify character, word and sentence counts, language patterns and key entities
6. **Text Summarization with T5-small:** We condense the content of publications into concise summaries that capture their essence, particularly for public benefit

NIHR BioResource for Translational Research – Facilitating Ground Breaking Genetic Research.

NIHR BioResource is funded by the National Institute for Health and Care Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.