Title

The Health Gym: An Open Platform with Health-Related Benchmark Problems for the Development of Reinforcement Learning Algorithms

Lead Applicant

Dr Sebastiano Barbieri

Details of proposal - team members and collaborators

Dr. Sebastiano Barbieri will be supported by Dr. Alan Blair (UNSW) regarding reinforcement learning, Prof. Louisa Jorm (UNSW) regarding data governance, privacy and ethics and Prof. Chris Dibben (University of Edinburgh) regarding the generation of synthetic medical data. Clinical expertise and access to relevant clinical datasets in antiretroviral therapy in HIV and sepsis management will be provided by A/Prof. Mark Polizzotto (The Kirby Institute) and Prof. Simon Finfer (The George Institute), respectively.

Alan Blair (UNSW) is a Senior Lecturer and Program Director in the School of Computer Science and Engineering. He has extensive experience in deep learning, reinforcement learning and adversarial training - with applications to strategic games, simulated environments, automated question answering, image synthesis and computer-generated art as well as machine learning and data analysis. He has recently been working on Bi-GAN models which are particularly relevant to the current project.

Louisa Jorm is Director of the Centre for Big Data Research in Health and Professor in the Faculty of Medicine, UNSW Sydney. She is an international leader in health ‘big data’ research and specifically in applying advanced analytic methods to large-scale routinely collected data, including hospital inpatient and medical and pharmaceutical claims data to create real-world evidence about the effectiveness of health care. She has earned more than $29M in competitive grant funding. Louisa is expert in privacy, governance and ethical issues relating to health data and has many high-level policy advisory roles, including membership of the Australian Health Ethics Committee, the national peak health ethics advisory body. She has led the development of major national infrastructure for data-intensive health and medical research, including the E-Research Institutional Cloud Architecture (ERICA) secure data laboratory. She is a high-profile advocate for more and better use of routinely collected health data for research.

Chris Dibben is Chair in Geography at The University of Edinburgh and researches population, health and place with a focus on poverty, deprivation and inequalities; evaluation of area-based initiatives; small area statistics; risk, vulnerability and hazards. Chris is the author of the popular synthpop package for R which supports creation of synthetic datasets that mimic the original observed data and preserve relationships between variables but do not contain any disclosive records. He contributed to work on the UK NHS health funding formula and on measuring health inequalities, for example developing the Health Poverty Index for the Department of Health and Information Centre (NHS). Other work includes the development of national deprivation indices across the UK and in South Africa and evaluations of government area-based initiatives.

Mark Polizzotto is an internationally recognized physician scientist with over 10 years experience in the conduct of clinical trials, with an emphasis on drug development studies in complex populations including people with HIV. He leads the Therapeutic and Vaccine Research Program, The Kirby Institute’s peak clinical trial program, providing leadership in the design and conduct of clinical trials and their operating procedures at all phases of drug development and in diverse settings including low and middle income countries. He leads several large international trials in therapeutics, including optimising therapy of HIV and its complications, currently accruing participants in over 20 countries. His work sits at the intersection of immunology, infectious diseases and cancer and his studies have been published in the highest ranked journals in each of those fields and have influenced treatment guidelines globally.

Simon Finfer is a Professorial Fellow in the Critical Care Division at The George Institute for Global Health, Adjunct Professor at UNSW and a practicing critical care physician. Simon is a member of the Global Sepsis Alliance Executive Board, and the council of the International Sepsis Forum. He founded and is Director of the Australian Sepsis Network and in 2017 initiated the formation of the Asia Pacific Sepsis Alliance. Simon's major research interest has been the design and conduct of large scale randomised controlled trials in critical care. Increasingly his research focusses on reducing the global burden of sepsis. He has obtained over $50M in research funding and authored or co-authored over 160 peer reviewed papers with 20% of those in the highest-ranking medical journals. He served as a guest editor for the New England Journal of Medicine from 2012 to 2014.

Details of proposal - vision, aims, influence on open research and evaluation plan

We will develop "The Health Gym", an open platform with healthcare-related benchmark problems, with a common interface and standardized data, for researchers to develop, test and compare reinforcement learning (RL) algorithms. RL is an area of artificial intelligence (AI) which centres on the problem of learning a behavioural "policy", a mapping from states or situations to actions, which maximizes cumulative long-term reward in an evolving, time-varying environment. The recent combination of RL with neural network modelling (deep RL) has led to algorithms with super-human performance in tasks from video games to complex board games, including go and chess. These successful applications of RL were greatly facilitated by the availability of online "benchmark problems": tasks in which an agent interacts with its environment according to specified rules and which provide a quick and easy way for the research community to develop, test, and compare RL algorithms (e.g. OpenAI Gym [4], DeepMind-Lab [5]). There is now huge interest in the application of RL algorithms to solve real world problems. In the healthcare domain, clinicians treating individuals with chronic disorders (e.g. epilepsy, mental illness, HIV infection) or with potentially life-threatening conditions (e.g. sepsis) often prescribe a series of treatments to maximize the chances of a favourable outcome. This generally requires modifying the duration, dose or type of treatment over time, and is challenging due to patient heterogeneity in response to treatment, potential relapse and side-effects. Clinicians often rely on clinical judgement and instinct, rather than formal evidence-based processes, to optimize sequences of treatments. Thus, there is vast potential for the application of RL algorithms for adaptive personalisation of treatment regimens, as shown by early research on optimizing antiretroviral therapy in HIV, radiotherapy planning in lung cancer, and the management of sepsis [6-8]. Nonetheless, some authors have highlighted the lack of reproducibility and potential for patient harm inherent in these methods [9]. In particular, recommendations made by RL algorithms may not be safe if the training data omit variables that influence clinical decision making, or if the effective sample size is small [10]. The main difficulty in developing robust RL algorithms for healthcare lies in the highly sensitive and confidential nature of clinical data, which often requires scientists to establish formal collaborations and execute extensive data use agreements before sharing data. One approach to overcome these barriers consists of generating synthetic data that closely resembles the original dataset but does not allow re-identification of individual patients and can therefore be freely distributed. Deep learning techniques such as (privacy-preserving [1]) generative adversarial networks (GANs) have recently been used to generate realistic medical time series [11]. The Health Gym will be developed in two steps. In Step 1 (months 1-6) we will use privacy-preserving GANs to generate synthetic clinical datasets. Initially, we will use the publicly available MIMIC-III data [2] and the EuResist data [3] to generate synthetic time series related to sepsis management in the intensive care unit and antiretroviral therapy in HIV. If the resultant GAN-generated datasets are realistic and the probability of disclosure in the (epsilon,delta) formulation of differential privacy is sufficiently small [12], we will extend the methods to similar clinical datasets held by our team. In Step 2 (months 7-12) we will create The Health Gym, an open platform with healthcare-related benchmark problems for testing RL algorithms. It will comprise a free and open source package for Python, and accompanying website with suggested "proof-of-concept" solutions and a wiki where researchers can share their solutions. We will also make the trained GANs and related software publicly accessible through an online software repository. At the end of the project, we will hold a two-day DataThon for teams of clinicians and data scientists, to help popularise The Health Gym platform and accelerate its use. How we will influence open research practices: Our project will influence open research practices in two main ways. Firstly, The Health Gym will be the first free and open source platform to make available healthcare-related benchmark problems for researchers to develop, test and compare RL algorithms. Such platforms have been critical to the very rapid and successful development of RL algorithms for gaming. The platform will provide researchers with open access to the highly detailed clinical data that are required to train robust healthcare RL algorithms, and which currently are extremely difficult to share. Secondly, the availability of our privacy-preserving GANs will facilitate the production of further synthetic clinical datasets for open sharing and reuse, with almost unlimited applications in medical research.

Monitoring and evaluation: We will adopt robust project management and agile working practices, including fortnightly showcases and retrospectives involving all team members, to deliver this project on time. Success indicators will be: Step 1: Synthetic time series data sets for both sepsis and antiretroviral therapy generated and validated, i.e. the correlation structure between variables and the causal effect of treatments is comparable to real data. Step 2: The Health Gym goes online with benchmark RL problems related to sepsis management and antiretroviral therapy, "proof-of-concept" solutions to these problems and a wiki. A two-day DataThon is held to popularise the Health Gym platform, with at least 5 teams participating.

Decision

Funded

Comment on decision from Wellcome

This was an interesting proposal from a strong team. The application was innovative and had clear and potentially wide-reaching impact

Title

PROM: Platform for Reusable Open Models for Predicting Antimicrobial Resistance

Lead Applicant

Dr Anshu Bhardwaj

Details of proposal - team members and collaborators

1. Roberto Toro is researcher at the Institut Pasteur in Paris, and research fellow at CRI, with expertise in neuroinformatics, bioinformatics and the development of platforms for scientific collaboration.
2. Marc Santolini, Research fellow at CRI, Paris, visiting researcher at the Barabasi Lab (Network Science Institute at Northeastern University, Boston) and research collaborator at Harvard Medical School, has expertise in AI, open science and network biology, will provide expertise on customization of the Galaxy platform with respect to metadata definitions and data integration.
3. Jonathan Grizou, Research Fellow, CRI, Paris. Research Affiliate, University of Glasgow. He has expertise in combining AI algorithms and robotics into the physical and life sciences. Previously co-founded a robotics start-up. He will provide expertise on integrating machine learning models on our platform.

The project PI is a member of the Galaxy Community. Started in 2005, the Galaxy Project (https://galaxyproject.org/) is a web-based scientific analysis platform used by scientists across the world to analyze large biomedical datasets with focus on making analyses accessible, reproducible, and simple so that they can be reused and extended (PMID: 29790989). Members of the this open source community will be engaged from time to time in developing the state-of-the-art platform.

Details of proposal - vision, aims, influence on open research and evaluation plan

(i) Aims: Given the increasing scourge of antimicrobial resistance (AMR) it is imperative that antibiotics are prescribed based on drug sensitivity profile (DST) of pathogens. Culture-based assays are currently the gold standard for drug susceptibility testing. However, recent studies have evaluated the combined approach of whole genome sequencing (WGS) followed by analytical tools as a preferred method to reduce time and cost of molecular diagnosis (PMID: 30886350; PMID: 29653190). A large number of these methods have been published and used to infer quantitative relationships between genomic variation and drug resistance phenotypes (PMID: 31047860, PMID: 30858411, PMID: 31182025, PMID: 30550564, PMID: 31106066, PMID: 30689732, PMID: 30333483, PMID: 31174603, PMID: 30514867). Nearly 30 prediction methods utilising WGS for predicting drug sensitivity profile are published every year. However, it is difficult to share these models on a common platform as there are no standards for depositing them. These models are currently available from code repositories, as services or as supplementary materials of publications. Lack of a standardised framework for sharing AMR models hampers their optimal use, limits their application on new datasets and makes it difficult to reuse them for composite models. Moreover, it requires expertise in data handling and machine learning to use these models. Given the pace at which genome sequencing is performed for various clinical isolates, it is imperative that a unified framework is designed to deposit and execute AMR models and datasets in a workflow system that makes it easier to reuse/reproduce the same. Therefore, the current proposal, aims at establishing an open workflow system that:

I) Is easy to use, not limited to machine learning / data handling experts

II) Overcomes challenges in customisation of input data formats and pre-processing both for reference and pan-genome datasets

III) Allows for hosting and community curation (crowdsourcing) of data on drug resistant determinants of priority pathogens based on FAIR principles

IV) Has provision of creating computational workflows and sharing them

V) Defines standards for building new models for interoperability and ease of reproducibility of findings with clearly defined metadata structure

VI) Offers a plug-n-play flexible environment for the end users to develop workflow systems for predicting AMR. Galaxy Workflow system is ideal for implementing this project as it allows for customization of all the above-mentioned features.

Target audiences: Researchers working in the field of AMR can use existing models on their datasets, contribute new models and datasets to PROM. Clinicians with sequence data can perform drug resistant profiling without getting into the hassles of installing and doing data pre-processing, which in most cases need expert/trained human resources.

Activities: PROM is expected to function as a centralised AMR models repository with existing machine learning models and input datasets (genome sequences, drug sensitivity and other annotation datasets) available as an open source platform for priority pathogens to begin with. Example packages will be provided for available models to test the system. New machine learning models will be built using the transfer learning methods. A network of research and clinician community will be engaged in testing the platform and contributing datasets/models to PROM.

(ii) The platform is expected to introduce good practices of data and model sharing as the models may be tested easily and the datasets are available in a ready to use format. This also provides an easy way for building reproducible workflows, for comparing different models on same input datasets and for selecting best models in any given context. Given the global challenge of AMR, it will also provide insights into how strain variability can lead to prediction of novel drug resistant determinants in different geographical settings. PROM will also help in avoiding reinventing the wheel for several steps involved in process.

(iii) 1. The first phase will deal with setting up platform for the critical pathogens (Acinetobacter baumannii; Pseudomonas aeruginosa; Enterobacteriaceae). This task will be divided into two steps: I) crowdsourcing data of these pathogens with respect to their drug sensitivity profile from literature as per EUCAST and CLSI standards. The PI has implemented several crowdsourcing projects and will assemble a team of 6-8 members for curating DST data on these pathogens. II) Simultaneously integration of existing machine learning packages will be done along with the input datasets as shared libraries. 2. Once the Galaxy platform is customised, benchmarking will be done with available models. 3. New predictive models will be developed in cases where datasets are not sufficient using transfer learning approaches. 4. Outreach will be performed with clinicians and research communities to provide training and hands-on workshops (at conferences like Epidemics/ICPIC). This will help in further customising the platform for the benefit of the research / clinical community and will also identify potential contributors. 5. APIs will be used to track the number of users, datasets/workflows to asses the usage/impact of PROM. 6. After finishing one cycle, several crowdsourcing activities will be launched simultaneously for generating data on other priority pathogens. 7. The outcome will be an open web-based platform with data on at least critical pathogens, their drug sensitivity profile, machine learning models and at least 3-5 workflows for predicting the drug resistance determinants with existing methods.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

The applicant opted not to share this information

Title

WebGEOMap JavaScript Leaflet & OpenLayers plugins using Standarised Detailed Hierarchical XML data-file

Lead Applicant

Dr Maksym Bondarenko

Details of proposal - team members and collaborators

Kerr David - SDI WorldPop, University of Southampton
Ves Nikolaos - SDI WorldPop, University of Southampton
David. K is a Computer Scientist with experience of a wide variety of the languages, tools and techniques of Software Engineering. For over 2 years, he has worked on research IT infrastructure development at the University of Southampton's SDI WorldPop - areas of expertise include Software Architectures and Engineering, System and Internet Security, Web development, Database design, Geospatial data management and Spatial Data Infrastructure, Semantic Web. David. K one of the developer the WorldPop Project website, portal and data management infrastructure. The development of the JavaScript plugin will be carried out by David Kerr, who has a significant amount of experience in web-based mapping, backend integration and computational numerical analysis of geospatial datasets, including the front-end development of the WorldPop data portal. This will involve the development of modules of the Leaflet and Openlayers plugins, to allow the users of the plugin to link their applications to the hierarchical file in order to analyse and parse its content;
Ves Nikolaos is a GIS Programmers. He gained experience through all stages of the project cycle, but over time he developed an increasing specialism for GIS, mapping and cartography, field survey logistics, data collection and data management. The development and structuring of the hierarchical file will be carried out by Ves Nikolaos, who has expertise in implementing server-side technologies within the group using open source tools, including the development of a REST API that allows researchers to access WorldPop data via their respective scripting languages. In addition to the development of the hierarchical data file, this developer will also contribute to the software that allows other formats to communicate with the data file (e.g. R/Python packages).Overall, David. K and Ves Nikolaos have a significant amount of experience in web-based collaboration systems and computational numerical analysis of complex systems. This experience in creating web-based centralised processing and database systems put the David K. and Nikolaos V. in a very good position to develop these WebGEOMap JavaScript Leaflet and OpenLayers plugins.

Details of proposal - vision, aims, influence on open research and evaluation plan

The use of geospatial datasets has increased recently, with the availability and use of these datasets in demographic and population health fields being no exception, particularly to achieve the UN Sustainable Development Goals from different perspectives. These data help decision-makers focus their analyses on different populations within countries to ensure the most vulnerable and isolated are highlighted when planning interventions. From a health perspective, these data can help analysts calculate the proportion of a population's access to a specific service. When considering a group within a population, such as children or pregnant women, these data become more powerful in assessing regions' access to facilities. The same data can also help with planning vaccination programmes or emergency preparedness. Whilst researchers are invariably adept at understanding and making use of geospatial data, there is still a need to disseminate their discoveries in a manner in which stakeholders not accustomed to geospatial data can understand the findings, enabling them to make effective decisions regarding these outputs. One contemporary method in which information extracted from analysis can be disseminated effectively to large audiences in a cost-effective manner is with web tools. Web-GIS (Web-based Geospatial Information Systems) is an online method of visualising data on a web map, providing interactivity to allow users to carry out simple analyses. Although web-developers have the tools to carry out these functions, there is still a need for them to undergo some training in front-end development and design in order to produce professional web-maps, and require knowledge in geographic concepts. The time and cost involved in some developments can dissuade some bodies from providing these visualisations, thus limiting the scope and efficacy of their findings. Leaflet and OpenLayers are two JavaScript web-mapping libraries that facilitate a range of geospatial functionality and visualisations. Being open-source, they allow the development of customised plugins to be hosted on their respective repositories. The WebGEOMap JavaScript Leaflet and OpenLayers plugins aim to alleviate some of the constraints on data providers by enabling those with limited web-development skills to access, visualise and disseminate a variety of geospatial data held on data providers' repositories. By including the plugin in their web pages, developers are able to link their applications to a hierarchical file held on the data providers' servers. The hierarchical file will allow data providers to specify options that can be linked to developers' plugins, allowing customisation of their applications with minimal configuration. The link between the user and the configuration file will be a URL that the user includes in their JavaScript file, specifying the options required for their application. To our knowledge, this is a novel approach in linking data and configurations between data providers, developers and front-end users. The advantage of such a framework will allow a uniform method for easily implementing an application linked to data. An example output of the customisation could be a choropleth map that shows a metric specific to a country, summarised at the subnational level. Upon clicking on the polygons, various additional information can be displayed in a pop-up window. This data will be requested from the provider, with minimal input required from the developer. WorldPop have developed a proof of concept for this tool (https://www.portal.worldpop.org/), and hope to consult potential users and stakeholders in low-income countries to provide further relevant functionality for the tool. Considering the impact that WorldPop data has had in providing health metrics in low-income countries, it is felt that this data could be further utilised this tool. Target users are health providers and decision makers in low-income countries, as well as their funders. In our experience, some stakeholders have minimal GIS experience, and such a tool would help to visualise different metrics quickly and effortlessly, thus helping to increase the use of geospatial data, and helping to improve the quality of life of citizens, and increase transparency and accountability. The plugin will be hosted on the Leaflet and OpenLayers repositories with full documentation of how to use it, along with a newly developed Standarised Detailed Hierarchical XML data-file (SDHXML).

The 3 main outcomes expected from the development of the new tool standard are:

The development of the newly-established data standard (Standardised Detailed Hierarchical XML) to link data-providers' data with developers. This will potentially contribute to academic advancement and development;

The development of reusable JavaScript web-mapping plugins in addition to complementary Python/R packages, enabling researchers to query data through the new data standard;

The project will be finalised with at least one publication in a computer science journal.

Version control of the project will be enabled by hosting the project on an open GitHub repository, to which, users will be able to submit amendment requests and trigger collaboration with the group. We aim to gather feedback on the project when we debut the Beta version of the software during the WorldPop Winter School workshop in late 2020. Following this period, final changes will be made to facilitate a long-term release of the product, which will include a fully-functional library for quality control.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This was an interesting proposal from a team with strong expertise. However, the application was overly vague, and the level of innovation and reach was felt to be limited.

Title

Open, reproducible analysis and reporting of data provenance for high-security health and administrative data

Lead Applicant

Dr Jessica Butler

Details of proposal - team members and collaborators

The people involved in this work are experienced collaborators with a range of expertise managing and using high-security health and administrative data. Co-development with data guardians, researchers, and policy makers is at the foundation of our way of working. This participatory approach allows the key users of the outputs to be actively involved in building them, thereby improving their uptake and impact. We have permission in place to begin this work in the NHS Data Safe Havens immediately.

Professor Corri Black, University of Aberdeen, Clinical Lead of NHS Grampian Data Safe Haven (DaSH)*, Director of Aberdeen Centre for Health Data Science, Associate Director of Health Data Research UK. Will provide supervision and expertise in data governance, secure data environments, health informatics methodology.

Dr Milan Markovic, University of Aberdeen Department of Computing ScienceWill co-lead development of the ontology model and building of software tools for provenance data collection.

Ms Carole Morris, NHS National Services Scotland Acting Head of Service, Director of Electronic Data Research Information Service (eDRIS)*Will provide expertise in documenting the extraction and preparation of health and administrative data for research, and provide expertise in secure data environments for the workshops.

Professor Nir Oren, University of Aberdeen, Head of Department of Computing Science Will provide expertise in development of ontology models and reasoning about trust in complex systems for the workshops.

Dr Magdalena Rzewuska, University of Aberdeen Health Service Research Unit. Will co-lead the workshops and provide expertise in multidisciplinary and participatory research as well as improvement and implementation evaluation methods.

Ms Katie Wilde, University of Aberdeen Technical Lead of NHS Grampian Data Safe Haven (DaSH)*Will supervise documentation of work-flows within the Data Safe Havens and provide expertise in secure data environments.

*DaSH and eDRIS are secure, accredited research facilities which provide access to health and administrative data for research. They support access to full electronic health records and other national datasets.

Details of proposal - vision, aims, influence on open research and evaluation plan

Aim

To co-design, pilot, and evaluate FAIR (findable, accessible, interoperable, reproducible) methods for tracking and reporting data provenance in high-security data research settings Background A wide variety of national-level health and administrative data are available for research, including hospital admissions, outpatient visits, prescribing, education, work and pension, and census data. Access is strictly governed because the data are sensitive and participants have not explicitly consented to their use. Data governance focusses on preserving confidentiality by strictly segregating the data and processing from use for research. Raw data, cross-dataset linkage, data extraction, cleaning, and anonymisation are done separately from research by diverse data custodians (NHS information services, local authorities, government departments and national records agencies). Researchers themselves access minimal, processed, anonymised datasets within secure environments (Figure 1 in Additional Information).   The opportunity cost of this strategy is loss of transparency about data origins, processing history, and an increased risk of undetected error propagation. Current procedures for capturing and reporting data provenance are fragmented across data custodians, often labour intensive, and rarely shared with researchers. The team involved in this project includes both data guardians and researchers who have discovered errors in high-security data that were difficult to detect due to lack of provenance reporting. We believe that the risks to research quality due to the opacity of data handling are as large as the risks of privacy breaches.   Objectives 1. Develop and test scalable automated methods of capturing data provenance within a high-security data research setting. 2. Co-develop tools to report provenance that are acceptable to data guardians and also meet guidelines for transparency and reproducibility.

Activities

Objective 1. Capturing Data Provenance 1A) Document common workflows for extracting and anonymising NHS data in Safe Havens 1B) Create a formal model describing the entities (e.g. datasets), activities (eg. anonymisation) and agents (e.g. data guardians) which generate research data 1C) Evaluate and refine the model with data guardians and specialist analysts NHS Data Safe Haven work-flows captured in Activity 1A will be used to make a formal provenance ontology model using the W3C recommendation for recording provenance, PROV-O (Activity 1B). The accuracy of the model will be evaluated in a two-hour workshop with 6-8 data guardians and NHS specialist analysts, where we will seek to identify and refine points where the model inadequately represents the process (Activity 1C). Success of these activities will be evaluated using three metrics: Does the provenance ontology model meet PROV-O standards? How fully does the model represent data processing of active Data Safe Haven research projects? Does the model accurately reflect analysts' experience processing high-security data?

Objective 2. Reporting Data Provenance 2A) Create software to automate the recording of provenance within Data Safe Haven workflows 2B) Generate provenance reports of varied detail for research projects in the Data Safe Haven 2C) Evaluate and refine these reports in a workshop with data guardians and researchers using high-security data We will create a software tool to capture provenance data using the ontology model (activity 2A). The tool will allow data analysts to generate structured/machine-readable provenance reports which can be filtered to provide a full provenance record for a dataset for use within the Safe Haven or a summarised version for public release (Activity 2B). We will run a two-hour workshop with 6-8 data guardians and researchers working with NHS data to evaluate example reports with varying detail and to refine the design (Activity 2C). Success of these activities will be evaluated by: How much of the NHS Data Safe Haven workflow can be captured? Does the full provenance report contain the necessary information to recreate the research dataset? Does the summary provenance report contain information that may not be released publicly? Does summary provenance report sufficiently describe the method used to create the research dataset according to FAIR standards?

Target audience The methods developed will be useable in all NHS Data Safe Havens and will scale to the parallel system used at Administrative Data Research sites. All research using national health, education, criminal, and tax data use these sites. The summary provenance reports are necessary for rigorous assessment of any research using these data. They will be useful to researchers, reviewers and funders.

Influencing Open Research Practice

Tools for transparent capture and reporting of data provenance will be a step-change in open practice in health data science. None currently exist. We will make all outputs open-source and publicly available (detailed below). We will implement the methodology on all future projects in NHS Grampian Safe Haven (lead by co-applicants Black and Wilde, with the aim to implement across all NHS Safe Havens sites (lead by co-applicant Morris). This work will also provide information that is needed to decide how to govern sensitive data. Making provenance data from these settings available allows researchers to study how storage and processing methods effect data quality. This type of assessment is necessary for deciding how to most responsibly govern the use of high-security data.

Decision

Funded

Comment on decision from Wellcome

This was an interesting proposal, from a strong team, aiming to address an important issue. The application was innovative and had clear and potentially wide-reaching impact.

Title

Open Synthesis: ensuring that systematic reviews are verifiable, repeatable and reusable

Lead Applicant

Dr Neal Haddaway

Details of proposal - team members and collaborators

Neal Haddaway, Senior Research Fellow at the Stockholm Environment Institute – Neal will lead the project, coordinating the group's activities, coordinating the hackathon, and leading the drafting of working papers. Neal will also represent the Collaboration for Environmental Evidence, an organisation that publishes Open Access environmental systematic reviews and is interested in making all aspects of these reviews (e.g. data and code) more Open.

Tamara Lotfi, Coordinator at the Global Evidence Synthesis Initiative (GESI) Secretariat, American University of Beirut – Tamara will co-convene the group with Neal, arranging webinars, group discussions and assigning tasks to the advisory group. She will make use of her connections across the GESI Network (http://www.gesiinitiative.com/gesi-network/details/GESI-Network-members) and its partners to solicit feedback from the broader community of systematic review experts (the Community of Practice [CoP] we intend to create). Tamara will ensure representation (along with other members) of low- and middle- income countries.

Martin Westgate, Research Fellow at Australian National University – Martin will support the coordination of the hackathon, having co-led 3 previous hackathons within the Evidence Synthesis Hackathon series (www.eshackathon.org).

Vivian Welch, Editor in Chief of the Campbell Collaboration – Vivian will represent the Campbell Collaboration, an organisation dedicated to publishing Open Access systematic reviews in the social sciences, and interested in implementing Open Synthesis strategies developed by this group.

Jordi Pardo Pardo, Governing Board of Cochrane and Managing Editor at Cochrane Musculoskeletal – Jordi will represent Cochrane, which publishes systematic reviews in health and is interested in making its reviews more Open by implementing strategies developed by this group. Adam Dunn, Associate Professor at the Centre for Health Informatics, Macquarie University – Adam will advise on informatics in relation to evidence synthesis, in particular in discussions regarding how meta-research (research on research) can be enhanced by using Open Synthesis principles. James Thomas, Professor at University College London and Director of the EPPI-Centre – James will represent the International Collaboration for the Automation of Systematic Reviews (ICASR). James will advise on the implications of Open Synthesis for systematic review management software and the role of software developers in supporting Open Synthesis, having led the development of EPPI-Reviewer review management software. Elie Akl, Associate Professor of Medicine at the American University of Beirut – Elie is a founding member of the Living Evidence Network, a group working to create guidance and methodology for continually updated systematic reviews. Elie will advise on the relationship between living systematic reviews and Open Synthesis principles, particularly with relation to priority setting.

Matthew Page, Research Fellow at Monash University – Matthew will represent the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Group, and will advise on methodological aspects of Open Synthesis related to transparency and repeatability.

Carina van Rooyen, Senior Researcher at the Africa Centre for Evidence at the University of Johannesburg – Carina will represent the African Centre for Evidence and the Africa Evidence Network, a continent-wide collaboration interested in evidence production and use in Africa. Carina will provide input from a lower- and middle- income country perspective: in resource-constrained environments especially there is a moral argument to be made for effective use of scare resources that can maximise good and avoid harm. Thus, Open Synthesis can be grounded in principles of equity, equality, accountability and transparency.

Details of proposal - vision, aims, influence on open research and evaluation plan

Systematic reviews (SRs) are vital for rigorous evidence-informed policy and providing essential feedback to improve underlying research. SRs generate huge quantities of data – including lists of relevant articles and their findings – yet these data are almost never released. Where data are available, they are not standardised. Some 3,000 SRs are published yearly, each requiring screening of c. 3,000 records, equating to 75m records examined annually. However, this data is typically not shared and is instead wasted. This stifles replication and updating of SRs, causing research waste.

Aims: This project has three core deliverables: 1) Community of Practice (CoP) for individuals/organisations interested in Open Synthesis; 2) a hackathon – a highly interactive workshop dedicated to understanding barriers and facilitators to achieving Open Synthesis and developing actions and recommendations for stakeholders; 3) a roadmap to Open Synthesis, including definitions and implementable advice and recommendations for Open practices in evidence synthesis. We will establish a CoP to define pathways to achieve Open Synthesis (i.e. Open Science principles applied to SRs), aiming to facilitate the verification, reuse, efficiency optimisation and automation of SRs through the application of FAIR (Findable, Accessible, Interoperable, and Reusable) principles to SR data. Open Synthesis aims to maximise openness and reusability of SRs, reducing waste from repeating tasks already conducted by other researchers. Open Science is not new, but its application to SRs (e.g. via the Systematic Review Data Repository; srdr.ahrq.gov) has, to date, been limited.

Target audience: The project will build a CoP of methodologists, journal editors, software developers and other stakeholders interested in the mechanisms of Open Synthesis. It will produce recommendations and guidance for the broader community of SRs commissioners, funders, authors, editors, and publishers. The project will rely on stakeholder engagement: the CoP will be involved throughout to co-design the project and its outputs. In particular, the hackathon (point 3, under "Activities") will be a highly interactive, co-produced event involving diverse stakeholders. Whilst the primary aim of this project is to benefit healthcare, we will engage with relevant stakeholders across disciplines (e.g. software engineers) to benefit from their knowledge and tools.

Activities: The CoP will facilitate Open Data/Methods adoption by defining data structures, types, storage, and minimum requirements. Since most reviewers use these tools, Openness can be substantially increased by developing standard, interoperable file formats (e.g. for citation screening decisions). This interoperability will allow the content of reviews to be rapidly reused in part or whole, without requiring manual data reformatting or repetition of effort. 1. We will assemble an Open Synthesis Working Group (comprised of leading experts in SRs across disciplines) to define Open Synthesis as widely accepted by various stakeholders. This will involve a discussion of 'how Open should evidence syntheses be?', 'which synthesis processes can be Open?', and 'what incentives and tools may support this transition and overcome barriers to Open Synthesis'? 2. We will develop recommendations to support Open Synthesis, including the production of standard data structures, data types and minimum requirements for the outputs of systematic reviews (e.g. lists of included studies with extracted data). 3. Finally, we will convene a highly interactive workshop in the style of the successful Evidence Synthesis Hackathon event (www.eshackathon.org) to facilitate discussions and develop tools (e.g. technology to transparently record information identified through web-based searches). We will host this workshop alongside the Cochrane Colloquium in Toronto in 2020.

Impact on Open practices: This project will make better use of current synthesis efforts and reduce research waste by supporting SR authors and publishers (e.g. Cochrane) with guidance, tools and incentives to be more Open. For example, through tools to archive machine-readable review data in publicly accessible repositories. Furthermore, verification and replication will strengthen the accountability and robustness of SRs to external criticism. The project has a high likelihood of impact, with support from 4 global SR organisations (GESI, Cochrane, Campbell, CEE) and strong connections to PRISMA reporting standards (used across c. 180 journals). Their involvement will allow the outputs to be effectively and efficiently integrated into workflows, rapidly changing practices (Cochrane has > 11,000 members and > 67,000 supporters, for example).

Monitoring and evaluation: We will use a range of metrics to monitor and evaluate success, including:

The Open Synthesis group: membership statistics, team satisfaction, progress towards project goals

Tools: number of hackathon tools developed, number of tool accesses/downloads in year 1, number of hackathon participants and their backgrounds, methods for recognising and rewarding Open Synthesis practices

Communication and engagement with the synthesis community: blog/commentary readership, article Altmetric scores

We will assess stakeholder group and broader community's acceptance of the project's recommended Open Synthesis practices using a combination of online surveys and minimum of 10 semi-structured key informant interviews (performed by NRH).

We will trial the hackathon tools and recommendations on a small number of case study systematic reviews to understand what factors facilitate or inhibit successful implementation.

We will use focus group discussions at relevant events to better understand users' perceptions of Open Synthesis. Since framing of reporting standards is instrumental in affecting uptake and success, feedback will help tailor communications. Finally, a dedicated cross-disciplinary website will be developed, and monitored for engagement/impact.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This was an interesting proposal from a strong team. However, the deliverables were not well-defined, and the potential impact was limited by the relatively narrow domain focus.

Title

Strengthening data fluency and a culture of evidence-based decision-making through the use of geospatial maps to improve health in Africa

Lead Applicant

Prof Simon Hay

Details of proposal - team members and collaborators

This work will be conducted in collaboration with African CDC and its five Regional Collaborating Centers (RCCs). The Africa CDC is a specialized technical institution of the African Union that serves as a platform for Member States to share knowledge, exchange lessons learned, and build capacity. Africa CDC supports all African countries to improve surveillance, emergency response, and prevention of infectious diseases. Its five RCCs are East, Central, West, North, and South, based in Kenya, Gabon, Nigeria, Egypt, and Zambia, respectively. The implementation of the project will take advantage of the already established collaboration between IHME, Africa CDC, and the RCCs.

Two key Africa CDC personnel – Dr. Raji Tajudeen MD, FWACP, MPH, Head, Public Health Institutes & Research, and Haftom Taame Desta, Epidemiologist – will participate in the project implementation. Their main role will be to enhance the implementation of the project through facilitating communications with key individuals in the RCCs. In addition, they will lead the capacity assessment and gap analysis in data fluency and use in these RCCs, as well as advise on follow-up activities.

Dr. Kebede Deribe, GBD and LBD Collaborator, Brighton and Sussex Medical School, University of Sussex, will provide geospatial expertise and act as a primary advisor and local liaison on LBD activities with Africa CDC, RCCs, and National Public Health Institutes (NPHIs) to achieve project deliverables and goals. Dr. Deribe is currently leading the 'Global Atlas of Podoconiosis' project and brings expertise in geospatial mapping that aligns well with the current project. Drawing on his broad experience in GIS, remote sensing, and spatial analysis, Dr. Deribe will build capacity on the use of geospatial evidence for informed health decision-making and advance open research through the use of LBD mapping throughout Africa. The Local Burden of Disease team at IHME includes an Engagement Manager and Engagement Officers, who will provide support for this project's activities. The Engagement Manager will manage overall engagement activities and collaboration with Africa CDC and Dr. Deribe to ensure alignment with broader project goals and that specific deliverables are met. The Manager will work with partners to promote the use of LBD estimates and tools to build awareness and trust in open research, embed best practices that support using data for decision-making, and improve LBD estimates for health. The Engagement Manager will also provide support for one or more of the following activities: oversee the provision of technical assistance to the Africa CDC Task Force on Data Management and Use, including advising on policy translation; participate in the Africa CDC 3rd convening of NPHIs; and support other LBD workshops and trainings for RCCs and NPHIs.The Engagement Officers will collaborate with in-country partners to successfully implement activities and ensure that project goals and deliverables are met. They will implement activities in each RCC region to address the gaps identified in the RCC assessment visit including conducting workshops and trainings; participating in the Africa CDC 3rd convening of NPHIs; and working with the RCC and NPHIs to build awareness of LBD open research tools and how to use them to identify and address national and regional health priorities.

Details of proposal - vision, aims, influence on open research and evaluation plan

Vision: This project envisages enhanced data fluency and a culture of evidence-based decision-making to improve health in Africa.

Aims The aims of this project are to: Build the capacity of NPHIs to use geospatial maps to improve evidence-based decision making; Strengthen geospatial data sharing by providing accessible resources and increasing awareness about LBD estimates and tools; Improve LBD health estimates to inform health policy and program planning and maximize impact throughout Africa.

Target audiences This project proposes to help expand the awareness of LBD open research resources and build trust in LBD geospatial maps and tools by providing access to the tools, geospatial expertise, and training to key health influencers and decision-makers in the Africa Center for Disease Control (Africa CDC) and its Regional Collaborating Centers (RCCs) in Southern Africa, East Africa, and West Africa, as well as related Ministries of Health, NPHIs, and regional governing bodies.

Activities: This project will enhance trust in published findings, facilitate use of existing data, and empower researchers and decision-makers by providing free maps and other tools for use in policy and planning to improve health throughout Africa. Specifically, we will work with the Africa CDC to build the capacity of its RCCs in Southern, East, and West Africa, and their associated NPHIs or similar organizations, to manage, exchange, and use the novel LBD maps and tools to guide health policy and planning decisions in each region.   Making these tools available and embedding best practices for using estimates in national and regional decision-making will help create a culture of data sharing, management, and use among NPHIs and maximize the impact of health interventions across Africa. Specifically, the requested funding will support the LBD team and Dr. Kebede Deribe to provide technical guidance, training, and support to successfully implement the activities outlined below with the goal of furthering the use of open research to improve health.   Conduct technical assessment visits to three Africa CDC RCCs in Southern, East, and West Africa to: Identify priority gaps in the use of data in the RCC and/or member NPHIs;   Introduce LBD and related open research tools to RCC member states;   Strengthen LBD estimates for each region;   Share best practices for accessing and using LBD estimates to strengthen the RCCs, NPHIs, and other collaborating data management centers.   Provide geospatial expertise and guidance to Africa CDC RCCs, and their associated NPHIs or related organizations, on the use of LBD open research estimates and tools to build awareness and trust in open research, embed best practices that support using data for decision-making, and improve LBD estimates. This may include one or more of the following activities: Support the Africa CDC Task Force on Data Management and Use, formed at the 2nd Convening of NPHIs in March 2019, to develop a framework to guide NPHIs in how to improve data fluency and create a culture of trust in the use of LBD results;   Participate as technical experts in the Africa CDC 3rd Convening of NPHIs in March 2020 to report on assessment visits and current LBD activities in support of RCCs and NPHIs' management, exchange and use of geospatial data for decision-making, and promote the use LBD open research tools to help identify and address national and regional health priorities;   Conduct workshops or trainings for RCCs and associated NPHIs to address gaps found in the initial data fluency assessments in the South, East, and West Africa regions.

Influence open research practices This proposal will influence open research practices by increasing data fluency and embedding LBD open tools and resources into the natural course of business of RCCs, NPHIs, Ministries of Health, and key stakeholders throughout Africa. Specifically, it will strengthen the capacity of RCCs and NPHIs to vet, interpret, and use data. It will also empower policy-makers within national and regional bodies to use evidence to inform health policy and program planning. By providing geospatial expertise, maps, and tools to RCCs and NPHIs across Africa, we will support a continent-wide effort to strengthen the management and use of open research to improve health.

Success indicators The success of this project will be monitored and evaluated based on the degree to which Africa CDC, RCCs, NPHIs, and other researchers, as well as decision-makers, are engaged with and using LBD estimates, tools, and processes. This includes: Providing geospatial data and vetting estimates to improve LBD estimates;   Target: Geospatial data provided from each of the three Africa CDC regions.   At least 10 new collaborators from each region signed up to participate in the LBD collaborator network which

reviews, vets, and comments on LBD estimates. Participating in webinars, workshops and other trainings; Target: One webinar or in-person workshop for each of the three Africa CDC regions. Building a culture of data fluency, exchange, and use within Africa CDC’s RCCs and/or NPHIs; Target: At least one example of LBD estimates and tools being used to inform health policy and planning decisions in each of the three Africa CDC regions.
Decision *Shortlisted, not funded*
Comment on decision from Wellcome *This application was from a good team committed to advancing openness. However, some elements of the application lacked clarity. The level of innovation and the potential impact of this proposal to transform health research through openness was felt to be limited.*

<u>Title</u>

TyphiNET – Unlocking public health lab data on travel-associated typhoid for sentinel surveillance

<u>Lead Applicant</u>

Prof Kathryn Holt

<u>Details of proposal - team members and collaborators</u>

1) Dr Zoe Dyson, London School of Hygiene & Tropical Medicine. Dr Dyson is an academic researcher specialising in genomic epidemiology of typhoid and other bacterial infections. Dr Dyson has been awarded salary funding via a Marie Curie fellowship to work on the TyphiNET project at LSHTM, under Prof Holt's mentorship. She will collate and analyse genome data, conduct statistical analyses testing the validity of using returning traveller data to predict pathogen populations circulating in endemic areas (including their antimicrobial resistance), and determine the requirements for the TyphiNET platform (data inputs, analytics to derive from the data, summary statistics/visualisations). If the Open Research Funding is awarded, she will interact closely with the contracted software developer and the public health labs, test the system, and prepare materials for & attend workshops and other engagement activities with public health labs, typhoid fever researchers and other stakeholder groups.

2) Prof Stephen Baker, University of Cambridge. Prof Baker is a microbiologist specialising in molecular biology of enteric pathogens, particularly the typhoid agents Salmonella Typhi and Paratyphi A. He has been awarded a WT Senior Fellowship to work on control strategies for Paratyphi A, which includes a substantial component investigating genomic epidemiology of the pathogen in Asia. He will contribute this unique data to the TyphiNET project, facilitating the first comparison of locally collected Paratyphi A genomic data from LMICs (including a study site at Christian Medical College, Vellore, India) with that generated in HIC PHLs from sentinel travellers. He will also, via his extensive collaborative networks in Asia and Africa, contribute to dissemination and engagement with the clinical and public health microbiology and infectious disease communities in LMICs.TyphiNET is initially supported by three exemplar Salmonella national reference labs who have agreed to share their routine genomic data and associated country-of-travel data for the 2-year duration of Dr Dyson's funded fellowship. If this Open Research proposal is funded they will also work with us to overcome the logistical hurdles and establish streamlined workflows to unlock and share their invaluable data on an ongoing basis for open research benefit; and will co-host workshops to encourage engagement with the platform by other clinical & public health microbiology labs and stakeholders and to promote data sharing for public health:

3) Dr Marie Chattaway, Gastrointestinal Bacteria Reference Unit, Public Health England

4) Dr Deborah Williamson, MDU Public Health Laboratory, Australia

5) Dr François-Xavier Weill, Enteric Bacterial Pathogens Unit, Institut Pasteur, France

<u>Details of proposal - vision, aims, influence on open research and evaluation plan</u>

(i) Our vision is to establish an open online portal that provides up-to-date genomic surveillance data on typhoid fever pathogens circulating in areas where the disease is endemic. Typhoid is a notifiable infection in UK, Australia and other high-income countries (HICs) where the disease is not endemic but rather is associated with travel to endemic countries. Most of these HICs have adopted (or are in the process of adopting) genomics for routine characterisation of typhoid agents in national public health labs (PHLs); most also conduct epidemiological follow-up of typhoid cases to establish the country-of-travel in which the infection was most likely acquired. At present these data are not routinely released (due to logistical/resourcing and privacy concerns), and are thus not integrated into the public global genomic typhoid framework where they could be used to inform empirical treatment and identify emergence and spread of antibiotic resistant strains. Hence we are currently not making full use of the genomic data generated regularly by PHLs. Why does surveillance data matter for typhoid? Genomic surveillance data is uniquely rich, simultaneously providing information on antimicrobial resistance (AMR), strain background, and local/regional transmission patterns. Clinical outcomes for typhoid are strongly influenced by effective antimicrobial therapy, which reduces mortality from ~10% to 1% and reduces the risk of onward transmission. AMR patterns for typhoid differ substantially by location, and clinical breakpoints for antimicrobials (especially ciprofloxacin, the WHO-recommended therapy) are poorly defined; hence geo-linked genomic data is particularly useful to inform empirical therapy. In addition, typhoid pathogen populations show strong genetic signals of geographical restriction; movement of strains across country borders is easy to detect from genome data, which can clearly identify the spread of new drug resistant strains and often can pinpoint the origin.   Why PHLs in non-endemic HICs? We are avid supporters of building local capacity for clinical microbiology, AMR surveillance and genomics in LMICs. However the barriers are substantial and such projects are long-term. In contrast, HIC PHLs labs provide a ready source of typhoid genomic surveillance data, and we propose it could be very low-cost to free this data and put it to immediate global public health benefit. Supporting this, our preliminary data (attached, which we will grow through the TyphiNET project) shows that routine genomic data from UK and Australian travellers are very strongly predictive of typhoid strain types and AMR frequencies, in 5 endemic countries studied so far.

Aim of the Open Research project: (1) Develop open-access protocols and workflows to streamline the secure flow of de-identified data from HIC PHLs, thereby reducing the cost of data sharing to the point where it is long-term sustainable; and   (2) Promote the open sharing of pathogen genome plus location-of-origin data by PHLs for public health benefit.

Target Audience: public health laboratories in HICs; public health and clinical practitioners managing typhoid in HICs and LMICs.

Activities: - Working groups within each PHL to identify and resolve local requirements and barriers to sharing data (Jan-Mar) - Analyses comparing traveller vs locally collected isolates, including Paratyphi A (Jan-Dec) - Software development (Apr-Sep) - Workshop 1: "Open Data and Public Health Microbiology" in Melbourne (Jul) - Workshop 2: "TyphiNET - Open Data for Typhoid" in London (Oct) - Evaluation, reporting (Nov-Dec)

(ii) Influencing Open Research practices in public health reference labs - This project focuses on typhoid, but can be considered an exemplar. The principle of harnessing routine genomic data from PHLs to serve as readily-accessible sentinel pathogen surveillance is potentially applicable to a wide range of notifiable infections that are commonly acquired by travellers (other foodborne/waterborne infections, vectorborne pathogens, AMR); for example Shigella (dysentery agents), which are priority targets for vaccine development in LMIC and frequently notified as travel-associated infections in HICs. - While we will work initially with the 3 named PHLs, we plan to reach out to many others, and envisage this project will prove the value of the concept and motivate others to participate and contribute. Different HICs have different travel patterns, so each PHL contributes to coverage of typhoid-endemic areas (see attachment; e.g. Pasteur data skews towards Francophone African countries; PHE data captures other parts of Africa and South Asia). - We envisage the project will provide useful data that demonstrates the public health utility of genomics for typhoid management and AMR surveillance in LMICs. This will help to support and motivate the development of in-country surveillance (project-driven or national programs), and to establish and normalise the culture of genomic data sharing for public health benefit beyond country borders.

(iii) Evaluation and success indicators will consider: (1) Is the data useful? (a) How well does traveller data predict locally obtained data (for both Typhi and Paratyphi A); and (b) Case studies demonstrating the value of the data, e.g. to resolve the most likely origin of a suspected travel case in a HIC, or to detect spread of emerging resistance phenotypes between endemic LMICs. (2) Are the initial 3 PHLs routinely sharing the data as planned? (3) Have additional PHLs been recruited to contribute data (in principle or practice)? (4) What additional barriers or risks have been identified?

Decision

Funded
Comment on decision from Wellcome *This application was from an impressive team, proposing to generate an important tool. The proposal had a strong potential to impact health research.*

Title

A search-engine inspired platform enabling open sharing and access of both data and results from all genomic studies

Lead Applicant

Dr John Lees

Details of proposal - team members and collaborators

Zamin Iqbal – Group leader (Computational Microbial Genomics), European Bioinformatics Institute (EBI). Dr Iqbal is a pioneer of large-scale sequence search methods – he and his group will advise and develop methods to index genome datasets, and contribute to their further development. His role at EBI will allow tighter integration with data storage services, helping with real-time incorporation of new data. This will also help support the continuation of this project beyond this proposal.

Nicholas Croucher – Senior lecturer (Bacterial Genomics), Imperial College London. Dr Croucher's extensive experience in analysis and curation of large genomic datasets will help us design data and analysis used to enrich search results. Additionally, ongoing links with public health researchers at both Public Health England (PHE) and Centers for Disease Control (CDC) will allow us to incorporate their specific requirements.

Lauren Cowley – Bioinformatics prize fellow, University of Bath. Dr Cowley has worked extensively with pathogen genome data in both public health and bioinformatics focused environments, including in low- and middle-income countries (LMICs). Her current research includes making routinely generated public health surveillance data more open. She will advise on user needs from a broad range of perspectives.

Stephen Bentley – Team leader (Genomics of Pneumonia and Meningitis), Wellcome Sanger Institute. Prof Bentley has generated some of the largest genomic studies to date. We will use these as specific examples while developing and testing the prototype of our tool. Both myself and Dr Croucher have worked extensively with Prof Bentley to curate and analyse the data he has generated, making these datasets ideal to start testing our approach. He will also provide user-perspective from his group, who routinely perform meta-analysis and re-analysis of published data.

Gerry Tonkin-Hill – PhD researcher (Department of Biostatistics), University of Oslo/University of Cambridge. Gerry's previous research has focused on analysis of genomes from malaria and other eukaryotic pathogens. He will contribute his experience with these larger and potentially more complex sequences to ensure our analysis and tool will be applicable when it is extended beyond bacteria and viruses. Additionally, Gerry has developed and maintains a number of successful software packages in R, and will help write a programmatic interface to our tool for 'power-users'.

Nicola Sugden – PhD researcher (Centre for the History of Science, Technology and Medicine), University of Manchester. Nicola is an expert on the history, philosophy, and sociology of science, technology and medicine. She has published a paper on the ethical implications of rapid genome sequencing (https://doi.org/10.1002/geo2.66) and has followed this up by considering the potential to de-anonymize public data, and signing rights over to private foundations (https://bit.ly/2JPcZCL). On this project, she will use her expertise to provide a sociological and ethical perspective on making sharing of pathogen genomic data more open and fair, and how to avoid privacy and data-ownership concerns.

Liam Shaw – Postdoctoral scientist (Modernising Medical Microbiology group), University of Oxford. Dr Shaw has previously worked as a consultant on antimicrobial resistance surveillance for a joint Wellcome/Open Data Institute project and was awarded a Wellcome Data Reuse Prize in 2019. He will use this experience of working with open research groups on genomic surveillance data to help shape the development of our tool. Dr Shaw's affiliation with the Modernising Medical Microbiology group at the University of Oxford brings further insight into the significant amount of publicly available genomic data they have generated. He will help understand what these datasets are best used for, and how high-quality analysis produced by the group can be included in our tool. Additionally, Dr Shaw coauthored the above paper on the ethics of sharing genomic data, so will also provide his perspective on open research in genomics.

Details of proposal - vision, aims, influence on open research and evaluation plan

Background We identified commonly encountered issues with the current publication and data sharing model (see attachment) which have led to a lack of openness in genomic research:

Data are often only available as sequencing reads, and are challenging to access efficiently. To transform this into biological insight requires highly specialised knowledge, and even those in the same field may be unable to use an unfamiliar dataset. This causes duplication of effort in assembling the data, and no guarantee of reproducibility. Many people, particularly in LMICs, lack the resources to carry out this computation. Analysis of genomic data is generally limited to what was initially included in a publication: constrained by the scope of the paper, arbitrary journal requirements, and may not even be open access. Although useful analysis and metadata are routinely produced for such studies, there is no easy or standard way for researchers to share anything other than sequence data itself. Sequence databases cannot serve all needs due to the diversity of biology, and research. Rules to ensure good quality uploads in one species may prevent use in another.

Aims & activities Innovative methods to index, rank and annotate sequences will be developed to solve these issues. We will adapt ideas used in web search engines, incorporating our biological knowledge, and engaging with our users. Key steps will be: 1) Searching data: We will first focus on pathogen data, adapting recently developed sequence indexing tools to work with traditional indexes on other data fields. Search engine-like 'crawlers' will be deployed to continuously update these indexes. Arbitrary annotations of samples from other databases will be linked using a NoSQL database. These details will be invisible to the user, who will simply enter search terms or sequence fragments to get results. 2) Ordering search results: Effective search strategies must be combined with an intelligent ranking system. We will develop a ranking method for results using metrics calculated from available metadata and analysis, quality of the sequence data and linked publications. Pre-defined ranking models and machine learning approaches which learn from the data will be compared. 3) Enhancing results with automated analysis: Our tool will include modules capable of adding further information to search results, to help users immediately evaluate sequence quality. Useful measures such as sequence length, number of genes/truncated, and k-mer based functional annotation will appear when hovering. At a population level, associations of variation with phenotypes will be reported. Further data will be scraped from publications, where possible. 4) Including researchers: As well as 'raw' reads, we will allow straightforward inclusion of processed data, such as assemblies and alignments, potentially saving hours of analysis per sample accessed. We will automatically generate flexible JSON data descriptors from arbitrary analysis results which include algorithm parameters and version. Sophisticated links which cannot exist in current databases will be possible, e.g. between sequence search results and visualisations with stable URLs. Users will be able to add annotations to datasets, such as suggesting potential reuses and rating results, which will also feedback into the ranking algorithm. 5) Outreach: We will deploy a pilot of the tool and identify potential early-adopters through our team, journals and twitter. We will host an in-person workshop/hackathon at the ABPHM conference, allowing us to watch how researchers actually use the tool, allowing them to discuss its advantages and shortcomings, and potentially add new features.

Benefits for open research:

Allowing researchers to provide structure-free information about their studies makes the barrier to open research considerably lower, encouraging sharing and data reuse in a way not at all facilitated by the current publishing model.

Increased democratisation of genomic data, opening research to a wide range of disciplines.

Much easier access for researchers in LMIC settings that may not have bioinformatic infrastructure or access to all scientific journals.

Users will be able to explore results starting from biology (sequences), rather than a complex path of publications, supplementary data, database and re-analysis of data which need to be manually linked one at a time.
Universality – DNA sequence is at the core, not organism- or population-specific quantities.
No changes are required to established database schemas, and no need for dedicated curators.
Continuous effort to integrating user feedback throughout the entire development process, rather than just at the end.
Monitoring success: This proposal will result in a pilot of our genomic search engine. We will be able to measure number of uses and unique users of this service over time, and we will also collect anonymised information from users who opt-in (proportion from genomic/non-genomic research, public health, LMIC users) to more specifically determine which groups we have successfully reached. Direct feedback will be possible through an issue tracker. Our focus groups will provide ongoing feedback on our success, and we will compare the changes in experience when completing a specific open data task both with and without our tool (attachment). Success in improving sharing in publications will be evaluated by continually measuring the number of journals and users using the prototype. Wellcome Open Research, eLife and Microbial Genomics would all be contacted due to their stated and proven commitments to open research.

Decision

Funded

Comment on decision from Wellcome

This was an ambitious application, from a strong team, proposing to generate an important tool. The proposal was innovative and had potentially wide-reaching impact.

Title

Practical steps to improve metadata quality and data sharing in circadian research

Lead Applicant

Prof Andrew Millar

Details of proposal - team members and collaborators

Dr Tomasz Zielinski (School of Biological Sciences; UoE Grade 8, lecturer-equivalent) is a senior research software developer, and the lead architect of BioDare2. His background in experimental research has helped our team to form close, working relationships with all the research groups for which we have created or adapted data management software. He will design and specify the new software features in BioDare2 that are necessary for the activities described in this proposal. He will supervise the implementation of the new features by a research software engineer funded by this award. Millar and Zielinski will supervise the analysis, presentation and publication of data from the project, with input from the team members below. This method of working has been the norm in several previous projects, where we have often worked with research software engineers from EPCC, Edinburgh's Advanced Computing centre.EPCC (formerly Edinburgh Parallel Computing Centre) is a major European facility with very broad expertise. EPCC hosts UK national facilities (e.g. the Archer supecomputing facility; EPSRC UK Data Store), with a staff comprising about 40 hardware engineers and 40 research software engineers.

Dr Niki Vermeulen (Senior Lecturer in Science, Technology and Innovation Studies, School of Social and Political Sciences, University of Edinburgh) is a social scientist specialising in science and innovation policy and the organisation of research, with an emphasis on scientific collaboration in the Life Sciences. Her recent research with Rosalind Attenborough (previously Publications Manager at PLoS Genetics) has focussed on Open practices in biomedicine. A PDRA under her supervision will be partly funded by this award, to design methodology and evaluation of user attitudes towards open research and data sharing. Dr. Vermeulen's salary is funded by SFC, not by this award.

Dr Andrew Romanowski (School of Biological Sciences) is a Grade 7 research data curator in the Millar group, with a background in molecular genetics and bioinformatics. He will score metadata quality and prepare the test examples that will train our users to rank the metadata in real data sets. Dr. Romanowski's salary is funded by research grants held by Professor Millar, not by this award.

Details of proposal - vision, aims, influence on open research and evaluation plan

(i) Vision. Accessible and reusable data are pillars of FAIR and Open Research but quality data deposition requires time and substantial effort. Linking sharing to data analysis can compensate for the time spent. However, quality sharing that considers future data consumers will require a change in research culture to create endogenous motivation for sharing rather than an external mandate. We will implement and evaluate low-cost, transferable methods to improve metadata quality and to influence attitudes to sharing, in BioDare2 (https://biodare2.ed.ac.uk/) the established data repository for circadian biology.

Target audience. Daily, circadian rhythms are a fundamental property of cellular regulation, acknowledged by a 2017 Nobel Prize. Circadian research touches many fields of biomedicine, and our users work from parasitology, to immunology, metabolism and mental health. Circadian timeseries over several days are costly to obtain, and mathematical analysis is required to measure the timing features in the data. Analysis of and access to timeseries data are therefore critical, across the diverse, rhythm research community. BioDare2 (https://biodare2.ed.ac.uk) is the only public, online repository for circadian timeseries. BioDare2 provides fast data analysis, summary statistics and attractive data visualizations in a modern web interface, all to ensure the best user experience (Figure 1). Access to BioDare2 is provided on condition that data will be made public. Thus Open data sharing is a "side effect of using our analysis tools and is not perceived as an additional burden. On the contrary, BioDare2 increases research productivity as its analysis methods are faster and easier to use than any alternative. This approach has successfully attracted data from users, gaining over 340 000 data timeseries in the two years since inception, with continuing fast growth (Figure 2).

Proposed Activities. Successful data re-use depends on good-quality metadata. We have empirical evidence that enforcing strict metadata standards can not only impede the uptake of our service but also triggers undesired behaviours, for example duplicating previous experimental descriptions or entering random characters as metadata. Human data curation is a possible solution but has not been achievable with current funding, a common problem for community resources. With over 500 active user sessions per month, we can perform controlled experiments on the effectiveness of alternative methods to influence user behaviour and promote reliable data sharing. BioDare2 is therefore uniquely positioned to allow practical, quantitative evaluation of these approaches: (1) Automatic metrics of metadata quality will be derived from online text analysis, completion of fields, and similarity to existing records. These provide immediate feedback to users, during data upload. (2) Sharing badge scheme: the user and their dataset will receive badges depending on the user's openness and the quality of their metadata, based on automatic metrics (1) or human "triage" (3). Badges could convey both positive and negative messages. (3). Human "triage" to rank metadata quality: users will be asked to score the quality of metadata, comparing fabricated descriptions and real content. Peer rankings contribute to many online communities, such as the successful, question-review process in the software resource Stackoverflow. Participation in the ranking process can educate users and enforce the community's views of good practice. (4) Functional rewards, where users who provide richer metadata or earlier data release gain access to new visualization or analysis methods. This approach can estimate the user's acceptable level of openness or metadata effort, in terms of the level of benefits required to engender each behaviour. (5) Questionnaires during the login process will evaluate user experiences and attitudes towards Open research at the start and end of the project, and can also provide information and some training.

(ii) Influencing Open practices. These measures directly improve metadata quality and Openness within BioDare2, reaching users across multiple biomedical fields. Moreover, our peer ranking process could lead towards community-based data curation in future. The proposed evaluation of various online techniques for changing research culture is unprecedented to our knowledge. The results will provide invaluable input to the Open strategies of many community resources that depend on volunteer contributions, far beyond the circadian community.

(iii) Evaluation is a major component of the proposal. Our targets are: Automated evaluation of metadata quality for all datasets. The quality of metadata will be measured using automatic metrics developed in Activity (1), validated by manual scoring. Automated feedback to users (Activity 2) and peer evaluation (Activity 3) will encourage culture change towards better metadata. Improved metadata quality on average, as defined by these metrics. In particular, we will eradicate the lowest-quality (meaningless or duplicated) descriptions of experimental data. Improved attitudes towards Open research; reduced perception of risks associated with Openness. User attitudes will be measures by questionnaires before and after the intervention. Improved Openness in practice, targeting voluntary early release of 20% of datasets (note that all data become Open by default). Metrics of early release, and dis-aggregated measures to identify which groups have engaged, will be automatically calculated. We expect increased re-use of the data but we do not expect sufficient examples to evaluate robustly within a 12-month project. High-quality evaluation requires that each activity is introduced independently to selected sub-groups of users. We will also monitor the impact of interventions on user satisfaction, adapting accordingly.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This was an interesting proposal from a strong team. However, the application would have been strengthened by the addition of a behavioural science specialist, and it was felt the level of impact could be limited.

Title

Randomization-based time-locking of study pre-registration

Lead Applicant

Prof Roy Mukamel

Details of proposal - team members and collaborators

Mr. Matan Mazor, University College LondonMr. Noam Mazor, Tel-Aviv University. Matan and Noam are the team members with whom the novel scheme for pre-registration was developed (see Mazor M., Mazor N. and Mukamel R. A novel tool for time-locking study plans to results published in Eur J Neurosci. 2019 May;49(9):1149-1156. doi: 10.1111/ejn.14278). In the current application Noam will work with the dedicated programmer, in the development of a pragmatic user-friendly web-based implementation of our scheme, and Matan will be responsible for developing a visualized guide-tool explaining the theoretical and pragmatic steps for using the pre-registration took.

Details of proposal - vision, aims, influence on open research and evaluation plan

Aim The global aim of the current proposal is to facilitate transparency in scientific reporting, and more specifically, to demarcate an objective line between hypothesis-testing and exploratory parts of scientific writing. We will do so by building an easy-to-use tool implementing a novel scheme we developed for pre-registration of study plans. Pre-registration will be objectively verifiable without the need for early external review. Our main target audience is the neuroimaging community although our tool may be applicable also in other domains. By providing the broad scientific community with such a tool, we aim to increase the practice of study pre-registration in support of our global aim. State-of-the-art Researchers face a high degree of freedom in choosing a particular analysis path over others. If not properly addressed, this may render reported statistical significance values susceptible to increased type-1 error due to multiple comparisons. This is especially pertinent when dealing with rich and complex data, such as neuroimaging data in which there are many analysis paths to choose from (for example choosing among various pre-processing steps, or various analysis parameters to use). One way to address this issue is to register study and analysis plans prior to data collection and thus avoid biasing the choice of a particular analysis path by the observed data. In common pre-registration schemes, authors upload any material pertaining to their study process to a server (e.g. https://osf.io/, or https://aspredicted.org/), and receive a time-stamp corresponding to the time of upload. Once study results are published, authors can use the time-stamp of the uploaded material as validation of their pre-registration. However, since time of upload and time of data acquisition are not locked, the time-stamp is not an objective means to verify that registration necessarily preceded data acquisition, rendering this scheme equivalent to traditional reports without pre-registration. Current proposal Recently, we developed a pre-registration scheme that time-locks the registration process to precede data acquisition and provides an objective measure of such time-locking. Our scheme, published in European Journal of Neuroscience (Mazor et al. 2019), is based on the use of cryptographic hash functions and as proof-of-concept, we demonstrated its use in an fMRI study. By facilitating valid and objective pre-registration, and increasing pre-registration rates not only in the field of fMRI but in other fields as well, I expect the deliverables of the current proposal to have a broad impact in mitigating low replicability rates due to poor separation between hypothesis-driven and exploratory parts in scientific reports. Deliverables We expect completion of the current proposal to result in two deliverables:

A user-friendly web-based platform that will accept a file of any format (e.g. Word, MPEG, or a zip file containing multiple files or directories), or a link to an online repository (e.g. on GitHub or osf.io), and return a randomized experimental design. Importantly, the randomized experimental design is unique to the uploaded files, thus time-locking them to precede data acquisition in a verfiable manner. The tool will be published in community journals (e.g. HBM, or NeuroImage), and will be freely available to the community by posting it on public websites (e.g. lab website, OSF, etc.)   A video tutorial explaining the theoretical framework of our pre-registration scheme, and also a more pragmatic tutorial explaining the various steps of using the software tool for study pre-registration.

Measures for evaluation of success There will be several quantifiable and objective measures that will allow us to evaluate the success of our proposed deliverables. Deliverable 1 (web-based software): The use of our scheme and tool will be assessed on two time-scales. Short time-scale (12 months): in the short-term, we will monitor the number of website entries and file uploads requesting unique design randomizations for time-locking as an indicator of the number of studies using our pre-registration scheme. Long time-scale (36 months): in the long term, we will monitor the number of papers citing and using our tools. This measure will provide a delayed estimate of usage. Deliverable 2 (video tutorial): We intend to publish an online video tutorial on a formal outlet (e.g. Journal of visualized experiments; JOVE) which provides a count of the number of views and various other metrics. As reference, we have previously published a video tutorial which received over half a million views (Ossmy, O., Mukamel, R. J. Vis. Exp. (127), e55965, doi:10.3791/55965 (2017)). A more subjective measure will be user feedback we expect to receive through linked web-forums and social media. As a preliminary indication of feasibility/success, we note that the Wellcome Centre for Human Neuroimaging at UCL has recently integrated the use of our scheme in their imaging pipeline and now provides it as an option for their neuroimaging community. In addition, we have been in contact with the Open Science Framework (OSF) team (Prof. Brian Nosak), who have expressed interest in our method and agreed to explore together options for incorporating it in the OSF website.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This was an interesting proposal about study pre-registration. However, the evaluation plan was not clearly described. The level of demand for this was unclear and hence the potential impact of this proposal to transform health research was felt to be limited.

Title

An open-access repository of bio-images to improve AI-based diagnosis of infectious diseases

Lead Applicant

Dr Elmer Llanos-Cuentas

Details of proposal - team members and collaborators

Development and Deployment:
Pierre G. Padilla-Huamantinco, Health Innovation Lab, Instituto de Medicina Tropical Alexander von Humboldt
Jose A. Zapana-García, Health Innovation Lab, Instituto de Medicina Tropical Alexander von Humboldt
Study Design and Clinical validation:
Fiorela Y. Alvarez-Romero, Malaria and Leishmaniasis Research Group, Instituto de Medicina Tropical Alexander von Humboldt
Gabriel Carrasco-Escobar, Health Innovation Lab, Instituto de Medicina Tropical Alexander von Humboldt
Sponsorship:
Richard Bowman, Bath Open Instrumentation Group, University of Bath
Julian Stirling, Bath Open Instrumentation Group, University of Bath
Sharada Mohanty, AIcrowd

Details of proposal - vision, aims, influence on open research and evaluation plan

(i) Vision: Our vision for this project is to build a web repository based on images collected by low-cost digital microscopy in order to provide clinical samples on neglected tropical diseases for Artificial Intelligence (AI)-based diagnostics. We will leverage advances in computer vision and open science hardware to set up open-source, low-cost and portable stations in healthcare facilities (laboratory units) where health personnel can easily read, label and upload samples as part of routine activities. At the same time, this provide an open image repository where the scientific community can use them through the platform for the development of disease diagnostics. (a) Aims: 1. We will construct and validate the OpenFlexure Microscope (OFM) in clinical settings. The OFM is an open-source and 3D printed microscope which includes a precise mechanical stage to move the samples and focus the optics(see Fig. 1 and Fig. 2). The software for controlling the microscope runs on an affordable Raspberry Pi computer and allows users to see a live preview of the microscope camera. We will use 3D printable files from OFM GitHub repository and build it according to the documentation. Then, OFM will be evaluated by the development team following the Mechanical and Optics performance tests published by Sharkey et. Al (2016). After functional tests, we will proceed to evaluate the resolution of the OFM with a positive resolution target and get some images from different specimens to be checked by a laboratory technician. 2. We will design and implement an open image repository based on Common Objects in Context (COCO) dataset format and DICOM Standard for Pathology. COCO is a large-scale object detection, segmentation, and captioning dataset which was designed as a new kind of dataset for computer vision research. DICOM Standard defines the protocols for exchanging information (interoperability) in medical imaging and makes it possible for the Pathology domain to be a part of the whole healthcare process. According to clinical guidelines and lab technician's expertise, we will define categories and dataset format to enable a practical collection of images and interaction with different visual interfaces. 3. We will design and implement a web platform for data mining based on the principles of Bioimages Informatics and Human-Centered Design (HCD). We will create a sitemap of the web platform and design the wireframes and mock-ups. They will be evaluated by a small group from our target audience and we will redesign the prototype based on the users' feedback. (b) Target audiences: 1. We will tailor our platform to direct use specifically by (1) Scientists (Global Health, Computer Science, Biomedical Engineering, etc.), (2) Health personnel (clinicians, technicians, pathologists, etc.) in limited-resource settings. 2. We will also target two specific audiences by disease area: (1) Malaria and (2) Leishmaniasis.

(ii) How your proposal will influence open research practices: Data size are growing from day to day in an increasingly connected world. Nowadays, the need to understand large, complex, information enriched data sets has increased in healthcare. The ability to extract useful knowledge hidden in these large amounts of data and to act on the knowledge is becoming increasingly important especially in early detection, diagnosis, and medical decision making. However, the research and training are conditioned to a few open access datasets. This situation limits advances in healthcare for patients and support for medical practitioners. The proposal intent of making our platform (hardware and software) available to the global research community. Users will find the documentation in a GitHub repo to set up these low-cost stations in health facilities and to collaborate to an open image dataset (crowdsourcing) by following protocols of sample preparation, digitization, and labeling. Users will be able to modify hardware component and to add other open-source microscopes (e.g. FlyPi). Besides the technology, users can also suggest protocols for new neglected tropical diseases or other kinds where microscopy is a gold standard for diagnosis. This community may become self-sustaining, building knowledge and new staff to improve health, and providing useful information (images) to help the next research generation.

(iii) How we will monitor and evaluate our proposal, included success factors: (Aim 1) Comparison between OFM and a conventional microscope. To evaluate OFM performance against standard microscopy, a laboratory technician will be tasked to observe specimens (parasites samples) with OFM and with a traditional microscope and to make species identification. (Aim 2) Evaluation of image dataset by experts. Images will be taken at two experimental research stations (a research lab and a public health facility). Experimental research stations offer the possibility of taking many images in a reduced amount of time. At this stage, specialists from the research team will check and clean the dataset removing poor quality images and fixing inconsistencies in data. (Aim 3) Evaluation of User Experience. We will recruit two groups of participants: UX experts (3 participants) and the target audience (6 participants in total – 3 per category). To evaluate the web platform, the participants will fill a standardized Usability Questionnaire, then they will join to focus groups to share their opinions about the UX platform.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This was an interesting application from a strong team. However, the relationship of the proposal to existing resources and initiatives was unclear and the outputs management plan would have benefited from more detail.

Title

FAIR and open multilingual clinical trials data in Wikidata and Wikipedia

Lead Applicant

Lane Rasberry

Details of proposal - team members and collaborators

Daniel Mietchen, Data Science Institute at the University of Virginia, Principle Investigator of the Scholia / WikiCite project to develop the Wikipedia / Wikimedia platform based interface for discovering and visualizing scholarly publications in a free and open system analogous to the popular but closed product Google Scholar. Dr. Mietchen's primary concern is the academic publications, whereas this proposal to Wellcome would integrate clinical trial data into this network and also localize the interface for non-English pilot languages relevant to the developing world. Another way to say this is that Dr. Mietchen operates a Wikipedia-based tool similar to PubMed and Google Scholar, and this project would collaborate with him to integrate ClinicalTrials.gov data into this and permit non-English language use.

Details of proposal - vision, aims, influence on open research and evaluation plan

Vision Our vision for this project is to increase public understanding and global discourse of medical research by making cataloging data on clinical trials much easier to access, query, and visualize in aggregate in English and 3 pilot underserved languages. Our aims are to enable the following: through publication in Wikidata, professionals in clinical research will have radically increased access to routine data about clinical trials, including from ClinicalTrials.gov and PubMed beyond conventional clinical research data, and for the benefit of the general public and humanities research, through Wikidata we will pilot access to previously inaccessible social ClinicalTrials.gov data including integration with geolocation data, grant and funding awards, corporate financing, and demographic data such as nationality, gender, and ethnicity after sharing the data, we will document accessibility options for all kinds of people, including citizen researchers, to use it. While the primary initial userbase will be people who already use ClinicalTrials.gov, we seek to make this data accessible and interesting to undergraduate students of all disciplines and pilot data accessibility in non-English languages including Hindi, Bengali, and Swahili languages.

Open practices Openness and FAIR data integration is a strength of this proposal because we originate in Wikidata as an activist community in favor of free licensing and machine readability. Beyond our routine practice of openness, we will document and publish our application of Wikidata values and practices as a model for others to emulate. Our publication venue is in Wikidata which has been the most popular, FAIR, and open cross-domain data repository in the world since at least 2015. This project starts with semi-structured open data in ClinicalTrials.gov which we will map to Wikidata, thereby making it highly structured, FAIR, and accessible in the Semantic Web and in multiple languages. Perhaps more significant than our making this data FAIR is our intent to document our process as a case study to demonstrate the ways in which the data was inaccessible and not FAIR before. Currently, many researchers see ClinicalTrials.gov to be FAIR and open because they compare it to conventional data management. This project will demonstrate how much more open this data can be and what networked integration can accomplish.

Monitoring: Monitoring is a strength of this proposal because the Wikimedia platform is digitally native for reporting an established suite of communication and impact metrics. Instead of developing any monitoring program, we will autogenerate the standard monitoring and metrics report which the Wikimedia platform offers and which includes data such as number of items edited, number of users who make contributions, and number of passive readers who consume the content. The original contribution which we will make regarding monitoring will be in good presentation of the Wikimedia results and documenting our use of the platform and its free and open features as a model for others to emulate. This project will publish its output into Wikidata, the structured data general reference repository which is part of the Wikimedia platform. Since its inception in 2001 Wikipedia and the Wikimedia platform have developed a culture and community where anyone can edit and a mix of humans, human-operated semi-automated tools, and automated bots monitor the millions of edits which thousands of users make every day in more than 100 languages. Our view is that in comparison to any other general interest data curation project, Wikidata provides the most openness, transparency, and reach. The best way to describe the monitoring plan for this project is to say that we will use the native Wikimedia platform monitoring suite of tools and products to collect and report metrics including count of edits; count of reported changes or conflicts; count of errors identified in the source dataset; count of comments; count of active reviewers and volunteer participant editors; and audience communication impact. This project has the strength of having a designed monitoring system in place which we will not change. Instead, we will make a model report collecting the metrics which are relevant to this project, and we will document how we collect those metrics from the Wikidata platform and how we interpret them, and we will create documentation for anyone else to post data for research production into Wikidata and monitor their own projects after our model. Success indicators include the following: integration of records from 80% of ClinicalTrials.gov trials into Wikidata, with each trial having an average of 10 structured data statements of fact translation of a limited vocabulary for queries and the web interface to make this data accessible in English, Hindi, Bengali, and Swahili published comments - endorsement or other feedback - from a diverse community of 100 Wikimedia editors publication of documentation for anyone to model this project and in advocacy of FAIR and open data

Decision

Funded

Comment on decision from Wellcome

This was an interesting proposal from a strong team. The application was innovative and had clear and potentially wide-reaching impact.

Title

d-Harmony: A data harmonization platform for pre-registering acquisition of open neuroimaging data and facilitating interaction between the data collector and the consumer.

Lead Applicant

Dr Ayan Sengupta

Details of proposal - team members and collaborators

Prof. Dr. Michael Hanke, Institute of Neuroscience and Medicine (INM-7), Juelich Research Center, Germany. Michael has developed popular open source software for neuroimaging analysis (neurodebian.net, pymvpa.org), and data management (datalad.org). He is a pioneer of the open data driven approach of neuroimaging (studyforrest.org). Michael will provide feedback on project design and contribute his expertise in software development. In particular, he will work on the integration of d-Harmony with the DataLad software for decentralized research data management.

Samir Das, Montreal Neurological Institute, Canada. Samir is the Associate Director of Technology at the McGill Centre for Integrative Neuroscience at the Montreal Neurological Institute (MNI). He is in charge of the infrastructure and technical aspects for several neuroimaging projects in Prof. Alan Evans' group at the Montreal Neurological Institute. He is the system architect for the LORIS platform and is the lead designer for MNI Open Science initiative. His role in the project is to help in software development, deployment of the system, project management and future integration of d-Harmony with LORIS and helping to reach out to a wider audience in the MR research world.

Details of proposal - vision, aims, influence on open research and evaluation plan

(i) Vision: The Problem Although data sharing in the neuroimaging community is more and more common, data acquisition and study design are typically confined to the lab acquiring the data and is not collaborative in any sense. This has led to incomplete or impaired utilization of the publicly shared dataset and has hindered the overall growth of open science initiatives. For example, Lab X is acquiring data from a patient population on a state-of-the art 7T MR scanner for a neuroimaging study consisting of functional MRI, high-resolution structural imaging and localizer scans. Concurrently Lab Y is planning to study high-resolution structural imaging and diffusion tractography from DTI scans from the same patient population. Because there was no open research platform for collaboration between Lab X and Lab Y before the data acquisition, the data being made publicly available by Lab X will be not be utilized and Lab Y will possibly have to recruit the patients again for just a 10 min long scan. This process is not only difficult and repetitive but also the procedure is a waste of resources and increases the overall carbon footprint of neuroimaging research. The current state of neuroimaging research is dire need of an open science collaborative platform for data acquisition.

Aim of the Proposed Solution As a solution to tackle the aforementioned problem, we suggest developing a web-based data harmonization platform for pre-registering acquisition of open neuroimaging data. On the website of d-Harmony the lab planning to acquire a dataset will be able to upload/enter details of their planned acquisition protocols, parameters, targeted participant demographics, and task details. Submitting these on d-Harmony will create an empty BIDS dataset including the metadata like task-description.json, dataset_description.json, participants.tsv etc., version controlled with DataLad/Git ready to be published as a 'pre-registered/planned' acquisition in a free and open neuroimaging data-sharing platform like OpenNeuro. In this way a planned acquisition would be made available to the neuroimaging community before the actual acquisition takes place and interested individuals would be able to send in request to add/modify scans to the planned protocols. If the requested changes can be accommodated by the data collector, corresponding metadata will be modified. After a certain period of time with iterations/modifications made to the planned protocol by this collaborative approach, the status of the dataset becomes 'in progress' until the entire acquisition is finished and the data made publicly available. In this way the main aim of the project to make neuroimaging data acquisition a collaborative approach for better utilization of open data could be achieved.

Target Audiences d-Harmony will be designed to target the entire open neuroimaging data sharing community by establishing a tighter connection between data collectors and prospective consumers. Using this platform, data consumers will be able to request modification to the data acquisition protocol, outlining potential value and enabling data collectors to consider an extension. Without this additional collaborative step, publicly shared data may not reach its full potential in re-utilization thus leading to waste of resource and time.

Activities We plan to host a session in Organization of Human Brain Mapping 2020 (conference hosted in Montreal) and ISMRM 2020 (hosted in Sydney) for introducing/popularizing the software to the open research community of neuroimaging. Along with that we plan to make video tutorials and prepare documentation for beginners to use this software. We also plan to organize training workshops for communicating our open research plans to potential users and collaborators and promote uptake and use of d-Harmony in the wider research community.

(ii) Influence on Open Research: Open research is a relatively new initiative and the neuroimaging community has taken important steps towards popularizing data sharing and data analysis software. But our idea of an open source software platform for collaborative data acquisition is a novel concept. This initiative would encourage more research labs to start collaborating even from the experimental design phase of their respective study. This leads to more efficient re-use of data and proper utilization of resources. In the beginning the software will be used for open research in neuroimaging but we are very hopeful that it would be easily translatable to related fields of Medical Imaging, genomic studies etc. We envision that large multi-site longitudinal studies contributing to databases like UK Biobank, ADNI etc. will be able to use our adapted services for collaborative data acquisition and open data publication.

(iii) Evaluation: The d-Harmony software platform will be free and open source. d-Harmony will be easy to deploy independently, to make it easy for institutions to harvest the benefits internally, where ethics and consent problems might be less severe regarding data sharing. d-Harmony will be developed in a fashion that it is easily integrable with already existing data procuring initiatives or longitudinal studies. It will also be evaluated whether there is an existing software platform that could be modified to implement the desired functionality, in order to minimize effort and facilitate sustainability. Next studyforrest.org acquisition will be proposed on this platform.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This proposal was from a good team, proposing to generate an important platform. However, the application would have benefitted from more information about the planned incentives to encourage participation and a more detailed evaluation plan.

Title

Accelerating the adoption of open research practices with a collective action platform for academics

Lead Applicant

Mr Cooper Smout

Details of proposal - team members and collaborators

Cooper Smout, B.Sc. Hons. I, is a Cognitive Neuroscience Ph.D. Candidate at the University of Queensland (Australia) and a Researcher in Training at the Institute for Globally Distributed Open Research and Education. Cooper founded this project in September 2018 and has since developed a preliminary support base through conference presentations, personal contact, the project website, Facebook and Twitter. Cooper funded the minimal viable product for this project and plans to work on the project full time after completing his Ph.D. later this year. Cooper will direct the project development and promote it throughout the community.

Claire Riss, is the Communications and Outreach Coordinator at the Center for Open Science (COS). Claire has worked extensively in the field of community engagement and will coordinate the COS Ambassador network, develop marketing and communication materials and promote the project to the open research and health research communities.

Brian Nosek, Ph.D. is a Professor of Psychology at the University of Virginia (USA), Co-founder and Executive Director of the Center for Open Science (COS). Professor Nosek will utilise his extensive experience coordinating open science projects to advise the project team and facilitate contact with relevant organisations.

Jonathan Tennant, Ph.D., is a Research Fellow at the Center for Research and Interdisciplinarity (France), an Affiliated Researcher at the Institute for Globally Distributed Open Research and Education, and the Founding Director of the Open Science MOOC. Dr Tennant previously served as Communications Director at ScienceOpen, has advocated extensively for open research in various formats and is well established in the open research community. He will advise the project team and assist with community outreach, communications and marketing.

Virginia Barbour, Ph.D. is a Professor at the Queensland University of Technology (Australia), Director of the Australasian Open Access Strategy Group, Advisor for the Declaration on Research Assessment, and Steering Advisor for Invest In Open Infrastructure. Professor Barbour was one of the three founding editors of PLOS Medicine and has served previously as Chair of the Committee on Publication Ethics, Editorial Director of PLOS Medicine, and Editorial Director of PLOS Medicine and Biology. Professor Barbour will advise the project team and facilitate contact with relevant health-related organisations and initiatives.

Alfred Garcia, M.Sc.IT, is a Software Engineer and Co-founder of Codi Cooperatiu SCCL (Spain). Alfred has 16 years of experience leading teams and developing a wide range of projects, including the minimal viable product for this project. He will lead the platform development and coordinate platform and database hosting, data security and data management.

Details of proposal - vision, aims, influence on open research and evaluation plan

1.1. Aims Intense competition within academia limits the uptake of open research practices. For example, researchers' concerns about subsequent publication opportunities can limit the sharing of preprints during outbreaks (Johansson et al., 2018). If a critical mass of public health researchers were to unanimously declare their intention to share preprints, however, publishers would be compelled to endorse preprints or risk alienating the research community. This strategy, known as 'collective action', could accelerate the adoption of open research practices in academia, while mitigating risks to vulnerable sectors of the community (e.g., early-career researchers and under-represented minorities). Although the internet is increasingly being used to coordinate conditional commitments to economic (e.g., Kickstarter), cultural (e.g., CollAction), and political action (e.g., the 'occupy' movement), no mechanism yet exists to organise collective action in the global research community. We propose to build a platform that hosts collective action campaigns in support of open research practices. Flagship campaigns will ask researchers to contribute to new preprint-review platforms (PREreview v2 and Rapid PREreview, a Wellcome Open Research funded project) or exclusively support journals that adopt particular publishing practices (Registered Reports, Open Peer Review, Green OA, Gold OA, Platinum OA; for details see http://tiny.cc/1hoybz). Researchers will login with ORCID and pledge to adopt each behaviour of interest, subject to there being a critical mass of support in their community. PREreview campaign thresholds will be defined using a fixed number of pledges (e.g., 1000), whereas publishing campaign thresholds will be defined using the 'impact' of the pledging cohort (e.g., 5%, 10%; see below for details) and personally selected by each researcher. After a support threshold has been reached, the cohort of pledging researchers will be publicized on the website and directed to carry out their pledge in unison.

1.2. Target audiences We have partnered with open research organisations (e.g., COS, UK Research Network, ASAPbio; full list here: https://github.com/FreeOurKnowledge/documentation/blob/master/partners.md) and will draw on these and other networks (e.g., DORA) to recruit 'ambassadors' for the project. Ambassadors will be trained (e.g., in COS webinars), provided with marketing materials (e.g., animated videos, badges, social media templates, presentation slides, stickers) and motivated (e.g. using scoreboards) to recruit users from two primary audiences: (1) researchers who already practice open research; and (2) researchers who do not yet practice open research. The first audience will be targeted using open research networks (e.g., mailing lists, social media) and materials that market our platform as a mechanism to enable stronger commitments to open research. The second audience will be targeted using health networks (e.g., Academy of Medical Sciences) and marketing materials that highlight the personal benefits of openness (e.g., increased citations, reclaimed funding). Ambassadors will insert slides into presentations at both open research and field-specific conferences. We will also sponsor conferences, partner with library networks (e.g., SPARC, LIBER, CAUL) to host outreach sessions at member institutions, petition health research schools to include marketing materials in staff training modules, and encourage peer-to-peer marketing. We will market the Rapid PREreview campaign at Public Health researchers to increase preprint use during outbreaks. The remaining campaigns will be marketed to all health research fields initially and subsequently targeted toward the sub-field with the greatest uptake. We anticipate rapid growth in Psychology due to awareness of the 'replicability crisis', the presence of established communities (e.g., the Society for the Improvement of Psychological Science) and the predominance of Psychology researchers in the COS ambassador network (34%).

1.3. Target activities The project will incorporate four major activities:

Beta-test and refine the prototype platform (found here: https://www.freeourknowledge.org/)

Develop marketing materials

Release platform

Promote campaigns

2. Influence We believe that this project will influence open research practices in the following ways:

Empower partially-open researchers to strengthen their commitment to open research practices

Increase awareness and educate researchers on the benefits of open research

Facilitate cooperation between open research initiatives

Complement 'top-down' mandates (e.g., Plan S) with 'bottom-up' support

Protect researchers made vulnerable by mandates

Generate data on levels of support for different open research practices

Pressure publishers and institutions to adapt their policies

Stimulate innovation and adoption of new technologies

3. Monitoring and evaluation Software will be developed using an 'agile' methodology across biweekly 'sprints'. The success of each sprint will be evaluated using the number and difficulty of tasks delivered, the number of commits to a production code branch, and the number of blocked or backlogged issues. Community support for campaigns will be evaluated using the number of

pledges (all campaigns) and/or the ‘impact’ of the pledging cohort (publishing campaigns). Impact will be operationalised as the relative proportion of citations that referenced articles authored by pledgers over the last 5 years (normalised by year and calculated separately for each research field using Dimensions.ai; for details see https://www.freeourknowledge.org/pages/about/). Campaign metrics will be broken down by research field and displayed on the website. Campaigns will be considered a ‘success’ if at least one pledge activation threshold is triggered in a health research field (i.e., 1000 pledges for PREreview campaigns, or 5% support for publishing campaigns). We also plan to analyse the pledge data along various dimensions of interest (e.g., researcher demographics) and prepare these findings for publication.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This was an ambitious application which showed a strong commitment to advancing openness throughout. However, there were some concerns about how the impact of the resource would be evaluated. The application would have benefited from more information about the long-term sustainability of the tool.

Title

afrimapr : facilitating the use of spatial data in African public health operations and policy with reusable R software building blocks.

Lead Applicant

Dr Andy South

Details of proposal - team members and collaborators

Andy South, LSTM. Andy will lead the project, contribute to code, documentation, and book writing and teach train-the-trainer workshops in Liverpool and Ethiopia.

Robin Lovelace, University of Leeds. Robin will contribute to code and book writing from his experience of leading collaborative code development and open-source book projects. He will ensure the quality and usability of software developed by the project, especially in relation to spatial data access and processing. Robin is active within the Free and Open Source Software for Geospatial (FOSS4G) and R communities. He recently led the development of the most comprehensive current R geospatial book (Geocomputation with R) as an open source book project (https://geocompr.robinlovelace.net/). He is lead developer on several R packages, notably stats19 and stplanr, which are being used internationally to access open data for policy making. He is Principal Investigator of the Urban Planning and Transport Health Assessment Tool (UPTHAT), which recently secured funding from the World Health Organisation, so is well placed to contribute to efforts later in the project to seek funding to continue afrimapr activities. Paula Moraga, University of Bath. Paula will contribute to code, documentation, and book writing from her recent experience developing health data related R packages and a book on geospatial health data. She will provide a statistician's perspective on our work to make data more available. Paula develops innovative statistical methods and open-source software for disease surveillance including R packages for spatio-temporal modeling (SpatialEpiApp), detection of clusters (DClusterm), and travel-related spread of disease (epiflows). Her work has directly informed strategic policy in reducing the burden of diseases such as malaria, leptospirosis, and cancer. Paula has taught statistics at several universities and has been invited to deliver training courses on disease mapping and R at international workshops. She is the author of the book 'Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny' (2019, Chapman & Hall/CRC).

Anelda van der Walt, Talarify, South Africa. Anelda will lead the engagement with African data communities, the development of contextualised and suitable free and open training materials for short workshops, and the trialing of this training materials in workshops in Africa. Feedback about the efficiency of the training materials will be collected and used to improve the materials. Between 2014 - 2019 Anelda led the African Carpentries initiative through which thousands of learners from 13 African countries were introduced to contemporary open science and data-science skills. Anelda helped coordinate 120 Data Carpentry workshops for African researchers including 10 train-the-trainer events yielding 50 qualified instructors and many more in the pipeline. Her hands-on experience in teaching novices foundational data science skills (including the use of R) will support the development of training materials matching the target learners' level of experience.

Margareth Gfrerer, Higher Education Strategy Center, Ethiopia. Margareth will coordinate training and train-the-trainer workshops in Ethiopia including identifying workshop participants. Margareth is posted by the German International Cooperation (GIZ) as an Education Scientist at the Higher Education Strategy Centre (HESC) within the Ministry of Science and Higher Education. She has worked closely with Anelda to develop the digital literacy of Ethiopian researchers by initiating data carpentry training, hackdays and competitions reaching 1.000+ researchers. Jointly they initiated the R-Ladies Addis Ababa Group in 2017 and can reach out to 10 R-Ladies Sister organisation at Ethiopian Universities.

Julie-Anne Tangena, LSTM. Julie-Anne will conduct a needs assessment for operational staff in Malawi and lead on trialling software and resources there. Julie-Anne is a public health entomologist with experience in both research and mosquito control operations. She will be working in Malawi as a part of a skills development fellowship to improve the use of evidence in mosquito control. Julie-Anne has extensive public health experience in Africa and Asia, including leading multi-disciplinary collaborations. She will be well placed in Malawi to investigate the challenges of software-use by public health workers and researchers.

Details of proposal - vision, aims, influence on open research and evaluation plan

Vision : Researchers and operational staff can easily make useful maps and applications for Africa from spatial scientific data using open-source R software. A wealth of health-relevant spatial data are available for Africa including disease model outputs, human population estimates, vector abundances and health site locations. Unfortunately, in-country health programs and researchers rarely benefit from them. This project will develop software building-blocks to facilitate the use of such spatial data, including the creation of web applications. We will use R - a top data-science language, ubiquitous within research and becoming more popular for operational programs. There is a growing data-science community in Africa with high potential to develop software tools to address local issues. The aim of this project is to support this group (and others) by developing these easily usable software building-blocks, which can be used to create tools relevant to local circumstances. We expect that researchers will be early adopters and can help promote use in operations and policy.

Activities : 1. Develop R package(s) in the open on Github based on existing team experience. 2. Invite early feedback from our networks in the public health and R communities. 3. Submit package(s) to rOpenSci and Journal of Open Source Software to receive constructive peer review. Submit to CRAN to make accessible to R users worldwide. 4. Develop training and train-the-trainer resources. 5. Trial software and training resources in UK, Malawi, Ethiopia, and South Africa. 6. Write an open book targeted at entry-level R users with step-by-step reproducible instructions for applying the R components to local use-cases. At the heart of the technical solution will be a modular R package or packages with re-usable components targeted at entry-level users. Designing for ease-of-use will be a priority. One of the key functionalities will be to improve access and use of African administrative boundary data, so that scientific datasets can be converted easily into maps, tables and applications that are useful for local decision-making. For example, code like "afrimapr(country='mali', adminlevel=2, detail='simple')" would plot and return a simplified map that could be displayed in a web application. A template 'shiny' web application will be included that can load any spatial surface and calculate summary statistics for pre-loaded or other administrative boundaries. Documentation will include instructions on how to copy and modify this template according to user needs. Early in the design process we will consider other functionality such as access to openstreetmap, health facilities, population estimates and satellite imagery. We will keep in mind the tension between usability and flexibility, and whether additional functionality should be located in other packages. The package components would make it relatively straightforward to create an application like the runner-up created by the lead applicant for the Wellcome 2019 Malaria data re-use prize. Our software and resources will catalyse the use of spatial research data in Africa by a broader audience. The software will also provide data producers more efficient and standardized means to increase the reach of their data. To achieve a large and positive impact we see two main challenges that this team is well placed to address. Firstly developing useful and usable software resources and secondly achieving a high adoption of those resources. We have a depth of experience both in creating software that is well used and in teaching and community development. Collectively the team also has strong networks in Africa, both at research and policy level. We will consider our project successful if the software components are adopted and used across Africa and elsewhere in operations and research. Adoption takes time, we expect clear indications by year-end but true success will only be apparent when these components or their successors are in wide use in five years and beyond. For monitoring and evaluation, the project can be split into three foci :

A. Code 1. Development will be in the open (under a permissive open license) on Github allowing others to contribute. We will monitor community feedback. 2. Early versions will be

reviewed within the group and tested on researchers within LSTM. 3. Submission for collaborative open peer review at rOpenSci (an initiative fostering an ecosystem of open-source tools for open science - the team has experience of submitting and reviewing code for rOpenSci). 4. Submission to CRAN. 5. Submission to JOSS (Journal of Open Source Software) - an open access journal for research software allowing for the software to be cited and tracked.
B. Training resources 1. Training resources will be trialled in LSTM on research staff and students. 2. Informal training trials will be conducted in Malawi, Ethiopia and South Africa. 3. Formal train-the-trainer sessions will be conducted in the UK and Ethiopia. 4. Local Ethiopian trainers will deliver sessions at 3 regional universities and results will be evaluated.
C. Adoption 1. Good choice of R package name(s) will allow internet searches to track code use. 2. CRAN downloads will be tracked. 3. JOSS citations (unlikely to appear within the timescale of the project) will allow future monitoring. Finally the project will be evaluated at the end for sustainability according to numbers of active contributors and prospects for further funding.

Decision

Funded

Comment on decision from Wellcome

This was an ambitious proposal from a good team with a clear set of deliverables. The proposal was innovative and had potentially wide-reaching impact.

Title

Increasing access to open source spatial demographic data using worldpopR and QGIS plug-in

Lead Applicant

Dr Andrew Tatem

Details of proposal - team members and collaborators

Dr. Natalia Tejedor Garavito (University of Southampton) is a Senior Enterprise Fellow and technical lead of the Geospatial Data Technical group at WorldPop. She has initiated, led, and contributed to multiple high impact research projects as part of the team and as an individual. She has a strong specialisation on the use and training on GIS and serves as an advisor on geospatial data analysis, providing accurate, reproducible, and freely available datasets. She will assist the project delivery and coordinate and carry out workshops. Dr. Alessandra Carioli (University of Southampton) is Research Fellow at the University of Southampton and affiliated with the Universitat Autonoma de Barcelona. She is an expert R user and shiny app developer. She contributes regularly to the R coding community and is an R lady member since 2017. She has over ten years of experience in the field of applied spatial statistics and population modeling and forecasting, having worked in the most renowned demographic research institutes in Europe. She is editor to a demographic online journal (Demotrends). Her role includes providing expertise on the building of the R package and data visualization shiny app from a demographic perspective.David Kerr (University of Southampton) is a Computer Scientist with experience of a wide variety of languages, tools and techniques of Software Engineering. For over 2 years, he has worked on research IT infrastructure development at the University of Southampton's WorldPop Spatial Data Infrastructure - areas of expertise include Software Architectures and Engineering, System and Internet Security, Web development, Database design, Geospatial data management, Spatial Data Infrastructure, Semantic Web and QGIS. He one of the developers the WorldPop Project website, portal and data management infrastructure. Hi role in this project will be to deliver the QGIS plug-in.

Details of proposal - vision, aims, influence on open research and evaluation plan

Vision: Enable non-technical experts to use WorldPop spatial demographic data and integrate it into workflows to support health research and decision-making.

Background: WorldPop (www.worldpop.org) provides a wide array of open access spatial demographic data, which are widely used by scientists and decision makers across the globe, particularly in low- and middle-income settings. These include geospatial datasets on population distributions, demographics and dynamics at 1x1km grid squares or finer, e.g. the mapping women of childbearing age, pregnancies and live births recently funded by the Wellcome Trust. These geospatial datasets have proven valuable to researchers and implementers in health and social research (> 10,000 citations, > 500,000 downloads), particularly in the field of health metrics and in tracking progress towards sustainable development goals, where the use subnational data is increasingly emphasised. However, the software required to analyse and visualise geospatial data, such as ArcGIS, often require expensive licenses, significantly limiting the range of data applications, particularly for low- and middle-income countries. This makes it difficult for researchers and health practitioners to use such geospatial-data, in addition to a general lack of expertise in many cases in ArcGIS and/or coding. In this context, R-Cran, Python and QGIS are free software and currently widely used among health data scientists, for their flexibility in data management, reproducibility of outputs and quality of data visualization. Therefore, our aim is to make geospatial demographic data easy and ready to use for a wide audience of health scientists and practitioners without a specific training in geospatial data, through a far-reaching range of data formats accessible via QGIS plug-ins, and R packages. Being able to easily acquire open access demographic data with relevance to policy planning, health metrics and epidemiology, would help researchers and practitioners to access and share knowledge, positively impacting scientific outputs and beyond.

Methods: 1. We will provide a vignette on how to format input data to be acquired through a QGIS plug-in. This would make it possible for any user to employ WorldPop data, integrate them with their own datasets as well as to access repositories from different providers (e.g. World Bank, Humanitarian Data Exchange, International Public Use Microdata Series, Demographic and Health Surveys). 2. We will create an R package (worldpopR) that can access existing datasets. Users will be able to query builder to analyse and visualize the datasets in different formats, relevant to their objectives. The analysis tools will include descriptive, zonal statistics and geospatial analysis that will help explore datasets. Moreover, the package will provide functions for plotting and mapping geospatial demographics at different geo-spatial resolutions: e.g., identify the total number of population/births/pregnancies/unvaccinated children at the district level and/or producing heat maps to classify areas of high and low density for multiple years and multiple countries at once. The data visualization will be included in a shiny GUI app designed in collaboration with end-users, and shared broadly through a dedicated website, together with reproducible code. These packages and shiny GUI app will be stored in open repositories (such as GitHub and on the WorldPop website) for easy access and updatability. End users will have the resources to recreate and tailor the app to their needs. 3. Additionally, the QGIS plug-in will provide a menu of mapping tools, where users can link datasets from online repositories and any other geospatial datasets to visualize datasets, apply multiple classification techniques and have the option to export them into a readable format to perform further downstream analysis. 4. We will pilot and evaluate our R packages and QGIS plug-in by carrying out at least one dedicated workshop in a low-and -middle income country, where we have contacts through our UNFPA partners with health ministers/practitioners. These partners are working to address targets for universal health coverage including sexual and reproductive health and rights, as well as improving women, children and adolescents' health. These newly created tools will substantially simplify and improve efforts to track programs and develop health metrics at subnational scales. Moreover, the testing of the tools in multiple other workshops will be feasible through the GRID3 program (grid3.org), where geospatial data support and capacity training is given to low-income nation governments and universities. We will also participate in at least two international conferences (e.g. ASTMH), where we will put together a symposium to present the applications as well as the outputs and receive feedback from scientists.

Monitoring success: We will be able to track the number of users downloading our applications through GitHub and the WorldPop website. There we will provide links, training materials and information regarding the packages and plug-in, as well as a feedback form to collect suggestions anonymously, and a dedicated user survey. We will get feedback from workshop participants to improve the package from both academic and policy perspectives, and ensure regular consultation throughout the design phase. Furthermore, dissemination of the products will be alongside geospatial capacity strengthening activities carried out within WorldPop in countries where we collaborate for existing projects such as GRID3 (grid3.org). This will enable us to obtain further feedback from National Statistics Offices, other ministries, scientists and UN country offices in low- and middle-income countries.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This was an interesting proposal from a strong team. However, the level of innovation and the potential impact of this proposal to transform health research through openness was considered limited.

Title

Recovering lost research: creating an open index of conference papers and presentations

Lead Applicant

Prof James Thomas

Details of proposal - team members and collaborators

ContentMine, a leading text and data mining not-for-profit organization, is our key partner in this project. Their mission is making knowledge and text mining technology available to every researcher in the world.ContentMine was founded after receiving social investment capital from Shuttleworth Foundation in 2014. ContentMine was Founded by Dr. Jenny Molloy molecular biologist and champion of open science and Cambridge Emeritus Professor Peter Murray-Rust, a well-known open science advocate, who developed the mining of scientific literature as a new form of research (https://en.wikipedia.org/wiki/Peter_Murray-Rust).. Her work focuses on the creation of an open and sustainable bioeconomy for the public good, where research, tools and systems are accessible to all. Additionally, Dr. Molloy is Co-founder of Biomakespace (https://biomake.space/home) a Cambridge community bio lab, and the Open Bioeconomy labs project (https://openbioeconomy.org/) in Ghana and Cameroon. For the last five years, ContentMine portfolio of projects and partners has grown, allowed them to become an strategic technology partner of text and data mining tools to both Higher Education and Knowledge based organisations.The values under which their mission is delivered are:

The liberation of scientific knowledge from scientific literature to make it useful for everybody.

To support an open community that uses and promotes content mining.

To create an open code, protocols and resources for mining.

To work with partners to create better tools and support their knowledge extraction.

Additionally, Contentmine is committed to advocacy and support an active community of early career researchers via its Fellowship Programme (http://contentmine.org/fellowship-programme/).

Cesar Gomez, Project Director, will lead the Text Mining team at ContentMine. Cesar is an engineer by training who became deeply interested in the role and impact of open science and open IP in research and innovation. He also Co-founder of Beneficial Bio a not-for-profit Biotech organization that aims to disrupt the life science industry and make it accessible to underserve markets. Cesar became Director of ContentMine in 2016, and has successfully managed all ContentMine's projects since then. He holds a MSc From UCL and an MBA from Judge Business school at Cambridge University.ContentMine's main role will be to process a large corpus of conference papers and presentations identified by UCL as part of the open index database by adapting its open access code called (AMI). The process will consist on the following steps:

To process the selected corpus of conference paper and presentations PDF's selected by UCL and convert them into a machine-readable format. This is achieved with an Open source code developed within AMI called (NORMA). A tool to convert a variety of inputs into normalized, tagged, XHTML (with embedded/linked SVG and PNG where appropriate). The initial emphasis is on scholarly publications but much of the technology is general. Norma can be built with maven3 and requires java 1.7 or greater. Link https://github.com/petermr/norma

To organize and classify each document into a Cproject. This tool was written from and primarily used by AMI and NORMA. It is written in Java and uses maven 3 to handle dependencies. This enables it to process multiple papers in a single run without overwriting files. It also keeps all the data from each paper together in its own CTree. This includes metadata about the paper, images that may have been extracted from the paper and supplementary files such as tables. Link https://github.com/petermr/cproject

Use parsing rules to identify each section of the document, i.e abstract, heading, conclusions, references and removing unnecessary information such as marketing and branding material. This is done through the use of optical character recognition (OCR) using GOCR.

Use AMI stack and Machine Learning frameworks to extract all valuable information including; authors, tittles, year, conference, abstract and any other information that may be relevant for the database. AMI provides a generic infrastructure where plugins can search, index or transform structured documents on a high-through basis Link: https://github.com/petermr/ami

To output the data into a User-Friendly environment (UI) and make it open and available to a wide audience by making all the code and resources free and available through a GitHub repository. Link : https://github.com/ContentMine.

Details of proposal - vision, aims, influence on open research and evaluation plan

(i) Project vision In many disciplines, thousands conference proceedings are either posted online in PDF files by their organising societies, or published as 'abstracts' articles in journals and their supplements. Individual paper and poster abstracts are thus not separately indexed, making it difficult to find them without manually checking hundreds of pages of conference proceedings by hand. Conference presentations are sometimes the only currently available record of a relevant research study; and if such studies are left out of reviews of the literature, then findings may not be genuinely representative of current knowledge. This status quo wastes time and resources in the manual checking of documents, can result in biased conclusions being drawn (and/ or distort the prioritisation of further research), and also risks wasting the effort invested in the original research. This project will directly address this issue by creating a semi-automated, open repository of conference abstracts. Establishing this repository will: reduce researcher time spent on identifying relevant abstracts; facilitate the reuse of existing evidence; and may, in time, be suitable for integration into large, open bibliographic databases such as Open Citation / Crossref.

(ia) Aims and objectives This project aims to make conference abstracts (e.g. papers, posters, symposia), which are currently 'hard to find' using conventional search techniques, openly accessible; so that they can more easily be identified and used. The project will achieve its aims by meeting the following specific objectives:

To identify relevant websites and conference proceedings;

To create a semi-automated workflow to index these; and

To publish the index online.

(ib) Target audiences The key target audiences for this work are: Researchers needing to identify research – e.g. for systematic reviews, grant applications or literature reviews; Academic societies and conference organisers – i.e. this project will ensure their work is more easily discoverable; and Those involved in ensuring that we have a complete archive of research – e.g. for Open Citation / CrossRef, this service promises to fill an important gap in current data sources. Downstream beneficiaries include: funders, who will see more impact for their investment; and wider society from more efficient evidence synthesis and more reliable summaries of current knowledge to inform decisions.

(ic) Activities Project activities will be split between UCL and ContentMine, adopting the following structure. Websites and documents that will be included in the development phase of the conference index will be identified. (UCL) The existing 'AMI' workflow for PDFs that ContentMine has previously developed will be adapted to ingest the files identified in (1), parse them, and output structured data containing the citation information (authors, title, year, conference, etc) along with the abstract for each paper or presentation contained in the PDF file. A web-based application with the following functionality will be developed and published (UCL): A searchable index of conference abstracts already ingested. Export to standard bibliographic formats. Upload of specific PDFs for processing (i.e. if a document has not previously been indexed). An API to permit integration of the index / associated functionality into other applications. Potential linking to related PubMed and Microsoft Academic records. A 'stretch' goal will be to link conference records to subsequent research publications and expose these links via the index. All source code will be published under an open source license as approved by the Open Source Initiative. The EPPI-Centre at UCL has already developed several online applications to support research knowledge curation, and we plan to host this new service on existing infrastructure.

(ii) How the proposed project will influence open research practices Although simple in its conception and operationalisation, the proposed project will have important impacts. It will save considerable time and effort on the part of researchers; reducing both bias in synthesised research findings and waste in research effort. It will also increase the discoverability of conference abstracts; highlighting their importance, and helping to ensure they can be more efficiently identified. This project can also influence future publication practices: many studies presented at conference are not subsequently published, leading to a biased published evidence base. Once an expectation that studies presented at conferences should be formally published becomes more established, researchers should be more incentivised to do so.

(iii) Monitoring and evaluation We will identify a range of conference abstract file formats in order to inform ContentMine development. Initial estimates suggest that between 50 and 200 will be needed to ensure we have a system with adequate generalisability. The adequacy of the 'training' sample will be monitored continuously, and additional examples identified and labelled if/when necessary. We will evaluate the accuracy of the automated extraction system against a manually curated gold standard dataset. While this is not a formal research study, we anticipate that this aspect of the work will be of interest to the wider community, and we plan to publish the results of this evaluation in Wellcome Open Research. Finally, we will evaluate user satisfaction in the system through an online survey and feedback forms built into the website.

Decision

Shortlisted, not funded

Comment on decision from Wellcome

This was an interesting proposal to create an open index of conference papers and presentations. However, the potential impact of this proposal to transform health research through openness was considered limited.

Title

Linking neuronal function to cell identity through novel whole-brain neurochemical datasets and comparative analyses

Lead Applicant

Dr Chintan Trivedi

Details of proposal - team members and collaborators

The following collaborations will be critical to the success of the proposed project:

Professor Stephen Wilson (Wellcome Trust Investigator, UCL, London, UK): Professor Wilson's group has significant experience in genetics and molecular biology of zebrafish. Professor Wilson's lab will provide the facilities and expertise for whole-brain fluorescence immunohistochemistry, in-situ hybridization and high-throughput microscopy to generate the core novel database for the project. His group will also provide expertise in anatomical annotation of the data.

Dr. Isaac Bianco (Wellcome Trust Henry Dale Fellow, UCL, London, UK): Dr. Bianco's lab routinely generates whole-brain functional and anatomical datasets and have considerable experience in brain morphing techniques and analysis of whole-brain datasets. Dr. Bianco's lab will provide test functional datasets, imaging and brain-registration expertise, and assist in the development and testing of the open-source analysis tool.

Dr. Jason Rihel (Wellcome Trust Investigator, UCL, London, UK): Dr. Jason Rihel's group specializes in studying the influence of neuropeptides and neuronal signaling on sleep behavior in larval zebrafish. His group also studies the influence of genetic mutations linked to Autism Spectrum Disorder on brain activity and behaviour in zebrafish. Dr. Rihel's lab will provide plasmids for several neuropeptide labels for the novel database and generate test data for validation of our analysis tool. Dr. Rihel's lab will also provide whole-brain maps for neurochemicals that his lab studies routinely.

Dr. Harold Burgess (National Institute of Child Health and Human Development, NIH, Bethesda, USA): Dr. Harold Burgess' lab has been instrumental in developing a brain atlas, the Zebrafish Brain Browser for the community. They have also pioneered computational techniques in the field for morphing whole-brain data on to standardized reference brains and for automated segmentation of anatomical regions across the larval fish brain. Dr. Burgess' lab will host our novel database on the publicly accessible brain browser. Dr. Burgess will also provide guidance on development of the open-source analysis tool.

Details of proposal - vision, aims, influence on open research and evaluation plan

Vision and aims: A fundamental goal of neuroscience is to understand how brain activity drives behaviour. Two aspects are critical to this goal: Mapping activity of neurons across the brain during behaviour and, determining whether each active neuron contributes to excitation, inhibition or modulation of activity. This entails mapping functional properties of neurons to their chemical identity i.e. what neurotransmitters and/or neuropeptides define each active neuron. To achieve this, the project aims to comprehensively characterize expression and distribution of neurochemicals in zebrafish and develop methods enabling analytical comparison between neuronal activity and neurochemistry for anatomically defined regions across the brain (Figure2). Larval zebrafish offer the unique opportunity to generate diverse whole-brain datasets at high spatio-temporal resolution, with high throughput and reproducibility across individuals. These datasets can be mapped onto standardized references due to advances in brain-registration techniques. Recently developed atlases (http://zbbrowser.com, https://engertlab.fas.harvard.edu/Z-Brain/home/, https://fishatlas.neuro.mpg.de/) host repositories visualizing transgene expression and neuronal morphology. However, these databases lack functional or neurochemical data, precluding mapping of neuronal function to a unique identity. Researchers in the field routinely generate brain-wide neuronal activity maps while larval zebrafish engage in different behaviours. The tremendous potential of mapping these functional datasets to specific identities in order to accelerate research through comparative analyses is largely unexploited. Our vision is to leverage statistically-validated, voxel-wise analyses

between diverse whole-brain datasets to empower researchers to link neuronal function to molecular identity. We aim to achieve this vision by providing the community with following solutions: Generation of a novel whole-brain database: We will generate a novel data repository (Figure2A,B) consisting of whole-brain labels for 75 molecules representing neurotransmitters, neuromodulators and neuropeptides. These molecules cover a broad spectrum of known neuronal signaling peptides/proteins and are conserved across vertebrates. We have developed customized fluorescence immunohistochemistry and in-situ hybridization protocols to allow high-throughput imaging of whole-brain samples (multiple brains/label) using light-sheet microscopy, followed by registration to a standardized reference brain. We will analyse co-expression patterns between all 75 labels to accelerate the goal of defining neuronal properties based on molecular identity. Comparative analysis of diverse whole-brain datasets: To lower the entry-barrier for researchers with limited expertise, we will develop a graphical interface to perform voxel-wise analyses of whole-brain datasets. The tool will facilitate integration of whole-brain neuronal activity with datasets representing additional features, thereby allowing the user to tie function to underlying neurochemistry, transgene expression and anatomy. For example, if a dataset identifies neurons with activity correlated to eye movements (Figure2C), the tool could readily answer these questions: (a) What proportion of these neurons are in which specific brain region? (b) Which neurotransmitter, modulator and/or peptide likely represents their molecular identity in each region? (c) Are there transgene expression patterns that overlap with these cells? Voxel-wise differential expression (T-statistic) and co-expression (co-localization coefficient) analyses will be performed by bootstrapping over hundreds of iterations, followed by statistical testing. The tool will generate whole-brain maps of test statistic for each voxel and significance maps for user-input p-value thresholds (Figure2D). By generating output spreadsheets for brain-region specific quantification of statistical maps, we will facilitate data-driven generation of novel insights and hypotheses. Our primary target audience is the zebrafish neuroscience community. Over 60 labs attended the Zebrafish Brain 2018 conference, and the field continues to grow rapidly. We will present tools and data created through this project at the Zebrafish Brain 2020 conference to encourage uptake. We will generate an online tutorial for the software tool with a linked forum to address user issues. We will organize a workshop at UCL in November 2020, and a linked webinar for optimizing acquisition, standardization and analysis of whole-brain datasets.
Open Research Practices: We will host the new set of comprehensive whole-brain data (novel database of 75 neurochemicals and existing repositories > 250 transgene maps) accessible through http://zebrafishucl.org/ and http://zbbrowser.com . Users will be able to submit new, diverse whole-brain datasets to keep expanding this online data repository. This will strengthen the adoption of open-research practices within the community and enhance discovery driven research, as these datasets will continue to be analysed in various ways. The code will be hosted on Bitbucket (https://bitbucket.org/) to enable modification by experts and community-driven improvement of the tool. This will also ensure automated credit assignment to the community developers for their efforts. The tool will be developed using PyViz (open-source visualization/analysis packages in Python https://pyviz.org/).

Monitoring and evaluation: As we progress through the project, the initial versions of our tool and expression maps will be released to our collaborators. This will ensure rapid bug-fixing and offer an opportunity to receive feedback on essential features to enhance user experience. We aim to release the fully operational version of the tool and novel database at the workshop in November 2020. The workshop and tutorial will allow us to inform the community about embedding our tool and database within existing open-source workflows for whole-brain registration and anatomical mapping, leading to broad adoption. Our success will depend on adoption by the community, which will be evaluated through following measures: (i) tool downloads (ii) dataset downloads (iii) source code downloads (iv) Bitbucket contributors (v) citations and (vi) new datasets uploaded

Decision

Shortlisted, not funded

<u>Comment on decision from Wellcome</u>

This was a high-quality proposal to make a valuable data resource. However, the level of innovation, as well as the potential impact of this proposal to transform health research through openness, was felt to be more limited.