**INTRODUCTION**

UK Biobank is seeking provision of an informatics platform to allow access and analysis of UK Biobank data, accessible to approved UK Biobank researchers from academia and industry. The first sections of this document set out an outline specification for the informatics platform (this is indicative rather than a full tender specification) and at the end of the document there are some specific questions which UK Biobank wishes to pose to interested parties.

The aims of the informatics platform are to:

* simplify and maximise access to UK Biobank data (with extensibility to other datasets, such as those managed by other cohorts);
* facilitate exploration of UK Biobank data and allow analyses of these data in-situ;
* integrate the various UK Biobank datasets to allow researchers to more easily understand disease mechanisms and the links between phenotypic, genetic, imaging and other -omics; and
* facilitate access to UK Biobank data, particularly for researchers in low economic development countries.

For the initial platform implementation, UK Biobank anticipates working with a single provider offering an integrated platform/compute/storage solution (albeit the provider of the platform and the provider of the compute/storage may be different entities). UK Biobank anticipates that it will make only one award for this initial platform implementation, but it may retain the option to elect to implement more than one solution.

It is recognised, however, that a long-term solution may involve both a number of different platform providers, and a number of different compute/storage infrastructure providers. Such an ecosystem would allow researchers to select the platform and compute/storage solution they use to work with UK Biobank data based upon considerations of familiarity, function, usability, support, and cost.

**Background**

UK Biobank is a large-scale prospective study of ~500,000 UK volunteers that aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart disease, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and dementia. UK Biobank is a de-facto open-access resource: there is a clear access test which is that the researcher has to be a bona fide researcher and the research has to be health-related and in the public interest, and all researchers and applications are subject to the same access criteria.

UK Biobank represents one of the most genetically characterised research cohorts in the world and has enjoyed substantial investment from external parties to create additional high value, large scale, complex datasets that would otherwise not have been brought into existence.

A current example is the Whole Exome Sequencing (WES) of all UK Biobank participants that is being undertaken by a consortium led by Regeneron. Existing UK Biobank access methods (via the ability to download data) have been used to make available exome data for the first 50,000 participants. A further tranche of 100,000 exomes are anticipated to be available for research access in early 2020, however, given the scale of data it becomes increasingly difficult to make these data available via download. Exome sequencing of the remaining 350,000 participants is now nearing completion and these data will be available in early 2021. As this will result in a dataset of over 1.5 Petabytes in size, transfer by download is no longer viable and this requires a platform access solution in order to enable the research community to access the data and to store it securely.

In addition, funding is now in place for a programme to complete Whole Genome Sequencing (WGS) for the entire UK Biobank cohort. This WGS programme alone will generate over 12.5 Petabytes of sequencing data and represents a significant investment by UK Government, Wellcome and industry (in the form of GSK, J&J, AstraZeneca and Amgen) in excess of £200M (<https://www.ukbiobank.ac.uk/2019/09/uk-biobank-leads-the-way-in-genetics-research-to-tackle-chronic-diseases/>). Similar to the WES data, these WGS data will need to be stored securely and made available to approved researchers within the research community.

Over the next five years, it is therefore anticipated that the UK Biobank resource will comprise over 15 Petabytes of data based on current scientific programmes. In the medium term the UK Biobank resource will likely be further enhanced (although there are no firm plans as yet in place) by various large-scale –omics assays (such as proteomics and/or metabolomics).

As such, UK Biobank requires a new approach to data storage and access, which allows researchers to bring their analyses to the data through implementation of an access and analysis informatics platform.

Such a platform will remove the need to download data in order to work with them, and will no longer require institutions to have their own dedicated infrastructure in order to download and analyse the data.

**HIGH-LEVEL REQUIREMENTS**

The following diagram outlines the key components, each of which are described in more detail in the subsequent text.



**Figure 1.** Key components of informatics platform (Box 1)

**1. Informatics Platform**

The platform required should allow researchers to analyse UK Biobank data ‘in situ’, combine it with their own data (e.g. summary statistics from another study), and collaborate with approved researchers within their own organisations (and/or other organisations). As an aside, the great majority of applications to UK Biobank are applications which involve more than one organisation. Core components of the platform include:

* **Storage Services** – to host and provide access to UK Biobank data (including WGS data) and storage of derived data and additional datasets imported from elsewhere;
* **Compute Services** – to meet the needs of complex processing;
* **Platform Support Services** – to facilitate secure access to Platform resources, including audit, charging and billing, and helpdesk services; and
* **Access and Analysis Services** – to provide functionality to support primary, secondary and tertiary analyses of genetic sequence data alongside UK Biobank data.

The platform is anticipated to be hosted on infrastructure (and UK Biobank is comfortable with the use of cloud infrastructure) that can deliver the necessary flexibility, scalability, and security. In particular, the platform should be as technology neutral as possible – so that it is compatible with other APIs and other compute/storage solutions.

The following diagram outlines the major functional components and subcomponents within the proposed platform, which are described in more detail in the subsequent text.



**Figure 2.** Platform Core Functionality (and existing UK Biobank Research Data Storage Services)

***Storage Services***

The Storage Services will need to provide hosting and access to data including:

* **UK Biobank Data.** The platform will hold a copy of publicly available UK Biobank data, including (but not limited to):
	+ phenotypic data;
	+ genetic data, including cohort-wide genotype, imputed genotype, exome sequence and WGS data;
	+ activity and imaging data; and
	+ linked electronic health record data.
* **UK Biobank Services for Data Curation, Ingestion, and Versioning.** UK Biobank will be responsible for managing the ingestion of new UK Biobank data and making it available to platform providers for incorporation. The UK Biobank data will be updated by UK Biobank at regular intervals (typically bi-annually) as new data become available (such as the inclusion of additional imaging data as further participants are scanned, or refreshes of linked electronic health record data).

The platform will support the rapid incorporation of meta-data and data objects via automated methods (such as access to a Web-based UK Biobank API).

The platform will support removal of participant-related data following automated notification from UK Biobank of any participant withdrawals (though the number of withdrawal requests remains very low).

* **Data Import and Export services.** Researchers will need to be able to upload or link to their own datasets or other publicly available datasets for which such linkage is technically feasible. All such access will be subject to appropriate levels of authentication and access control. Researchers will also need the ability to be able to export certain UK Biobank data and derived data from the platform for further analysis using their own infrastructure.

***Platform Support Services***

The Platform Support Services will need to include:

* **Authentication and Authorisation**. The platform will include authentication and authorisation controls to enable researchers to implement appropriate levels of access control (so that only approved researchers are able to access the data), and will build upon the capabilities of the underlying service to ensure data integrity and availability. Access to the platform will be restricted to UK Biobank approved researchers, using the same credentials they use to access UK Biobank’s Access Management System (AMS) controlled via a real-time Web-API provided by UKB.
* **Security and Compliance**. The platform will be able to conform to the provisions of the European GDPR, particularly in relation to the ability to allocate where data are stored, and have ISO27001 certification or equivalent.
* **Audit**. An audit trail will need to be maintained of platform activity, including data access. Audit data will be retained for a period of 12 months, and will be reviewed regularly to identify anomalous activity and as necessary to investigate suspected instances of unauthorised data access or data leak.
* **Charging and Billing**. Costs associated with compute utilisation, storage of additional data (whether imported from external sources or derived from analyses undertaken on the platform), and data ingress/egress will be borne by individual research projects. This Charging and Billing component for platform access and consumption of compute and storage resources will provide utilisation reporting and mechanisms to allow these incremental charges to be passed through to the researcher (UK Biobank will continue to manage the application process and the basic access fees).
* **Data Management**. UK Biobank will continue to use its existing schema for all data made available through the platform. The platform will need to provide project workspaces, each of which will maintain a pseudonymisation mask (such that data can only be viewed using project-specific pseudonymised identifiers that have been issued under the researchers’ approved UK Biobank Access application). Researchers working across multiple projects must only be able to see the appropriate pseudonymisation scheme within the workspace for each project.
* **Helpdesk**. The core facilities of the platform (including the administrative tools, the analytic tools and workflows available natively, and the facilities for deploying additional tools, workflows, and integrations) will be supported by comprehensive documentation, supplemented by online tutorials and videos where relevant. A helpdesk will need to be provided, supporting interaction via email, contact via webchat or telephone during (as a minimum) normal UK business hours. There are two components to this helpdesk service: one which is available to UK Biobank, to manage the platform and upload UK Biobank data, and the other which is available to all research users of the platform, so that their questions, queries and service issues can be addressed.

***Compute Services***

The platform may be built on widely available cloud infrastructure and will offer access to a range of compute capabilities. These will need to include scalable instances of virtualised general purpose compute, available in different configurations of core and memory size. Depending upon the scope of services that UK Biobank is able to procure, this may include access to specialised compute including, for example, GPU-enabled nodes.

***Access and Analysis Services***

The Access and Analysis services will need to include functional components that provide:

* **Processing**. The platform will provide access to a range of utilities, tools and workflows. Researchers will be able to run analysis processes either by invoking them from the user interface, or by invoking them programmatically using APIs supported by the platform.
* **Automation and Workflow**. The platform will support a workflow language (for example, Common Workflow Language) for pipeline definition and automation.
* **Collaboration**. The platform will allow project specific utilities, tools, workflows, and datasets to be shared with other collaborators and/or other UK Biobank approved projects, subject to appropriate levels of authentication and authorisation.
* **Tools Repository**. The utilities, tools, and workflows available as part of the core Platform will be described in and accessible through a searchable repository. Tools and workflows developed by users of the platform can be made available as part of a searchable catalogue for reuse by others.
* **Workspaces**. Individual research projects (i.e. each UK Biobank Access application) will have their own defined project area within the platform in which to store any data derived as part of their research project, additional data that they have uploaded, and/or project specific workflows that they have created or uploaded. Objects held within each project area will not be made visible or available to other researchers, but collaboration capabilities will allow controlled sharing of algorithms and workflows with other research projects.
* **Containers**. Researchers will be able to deploy their own tools to the platform (subject to appropriate licensing and technical compatibility with the platform). To facilitate this, the platform will support deployment of containers using technologies such as Docker.
* **Interactive Analysis and Visualisation**. The platform will provide the ability for end users to interactively explore, analyse and visualise the data, using tools including:
	+ **User interface tools**. The platform will support a variety of user interfaces based upon standard bioinformatics research tools such as Jupyter Notebooks (with support for example, for Python) and R-Shiny applications.
	+ **Cohort Browser**. The platform will provide a capability to support data browsing and sub-cohort definition. It will be possible to define multiple sub-cohorts, save their definitions, and share those definitions with other approved projects in a controlled manner.

**Other Components (outwith the direct scope of the platform)**

For completeness, a description of the two other boxes outlined Figure 1 are included below.

**2. Support & Service Delivery Team**

UK Biobank will maintain a support and service delivery team, including bioinformaticians, data analysts, and developers to:

* provide direction of ongoing development of further functionality;
* assist with incorporating existing and new UK Biobank data into the platform efficiently;
* undertake QA/QC of data returned from external Access projects;
* enhance the value of these data to the research community;
* advise researchers on how best to use the capabilities that the platform provides;
* facilitate linkage to other relevant datasets; and
* work to harmonise UK Biobank data with emerging standards (e.g. OMOP).

On the assumption that the informatics platform will be provided as a managed service, the 3rd party service delivery team will need to work closely with (and complement) UK Biobank’s internal support and service delivery team.

**3. Extension of existing UK Biobank infrastructure services**

UK Biobank has existing data storage and access services (the UK Biobank Research Data Storage Services) which will be maintained alongside the informatics platform provisioned:

* UK Biobank Data Showcase – an on-premise system for access to existing UK Biobank data, together with data exploration and access services (via the ability to download these data using UK Biobank’s provided utilities).
* European Genome Archive (EGA) – a service hosted by the European Bioinformatics Institute, Cambridge (UK) for the secure storage and access to large-scale genomic datasets.

**UK BIOBANK QUESTIONS FOR PLATFORM PROVIDERS**

**When responding to the questions below, please do not provide any commercially sensitive information such as pricing. UK Biobank will make available a synopsis of the feedback received from responses to these questions and any video conferencing meetings, without attributing any response to any particular potential supplier.**

**When responding to the questions, please indicate whether your response is for delivery of services by yourself or whether you would typically engage others in your supply chain or joint venturing or be a sub-contractor etc. for any aspect.**

**1. Functional capabilities**

UK Biobank is seeking to procure a data access and analysis platform that will allow researchers to work with UK Biobank data in-situ rather than requiring them to download these data. The key functional components of such a platform include: a web portal providing controlled access to dedicated project area(s); a cohort browser to explore the data available and to select a sub-cohort of interest for detailed analysis; support for undertaking such analyses using interactive tools and batch pipelines provided by the platform; and the ability to use the platform to extend those analyses using additional tools and pipelines of their own.

* Please describe which functional capabilities are available within your platform.
* Do you plan to implement any capabilities not currently available; if so, when?
* What is the functional scope of your platform cohort browser (if applicable)?
* What interactive tools could your platform support?
* How could your platform support researchers deploying their own tools or pipelines?

**2. Scope of data**

The UK Biobank data include both structured (tabular/relational) data and bulk (BLOB) data (such as MRI images and genetic sequences) for participants in the study. The structured data currently includes over ~25,000 individual attributes for each participant including physical measures, questionnaire responses, imaging derived measures, and other phenotypic and demographic information; and linked healthcare records including secondary healthcare records for the whole cohort and primary healthcare records for ~40% of the cohort. The bulk data include exome and whole genome sequences (CRAMs, VCFs), genotyping array data, activity monitoring and imaging and are expected to grow to some ~15 petabytes over the next 5 years. The data types and total volume will increase in future (for example, if funding becomes available for e.g. proteomic and/or metabolomic studies).

* What range and scale of the data described could your platform support?
* What tools could you provide (or support) for analysis of structured data?
* What tools and pipelines could you provide (or support) for analysis of genomic data?
* What tools and pipelines could you provide (or support) for analysis of other data types (e.g. imaging, proteomics)?
* What plans could you have to provide (or support) tools not currently covered?

**3. Infrastructure**

UK Biobank considers that the compute/storage capacity of the platform will need to be global and flexible.

* On what infrastructure could you propose to implement your platform?
* In what circumstances could you propose a public cloud implementation?
* In what circumstances could you propose a private cloud implementation and/or dedicated physical infrastructure?
* On what basis could you support (and how could you charge for) platform data ingress/egress? \* please do not provide actual pricing information.

**4. Managed Service**

UK Biobank wishes to procure the platform as a managed service, whereby the contracted party will be responsible for the platform implementation, operation, ongoing enhancement and support.

* Could you offer your platform as a fully managed service?
* What are the main elements of the service you could provide to support your platform?
* How do you envisage support could be provided for researchers being shared between your team(s) and UK Biobank (e.g. first and second line helpdesk)?

**5. Billing and Charging**

UK Biobank would expect to bear the initial cost of the core platform provision (i.e. platform software, compute, storage of the UK Biobank data, and basic service support) but not the costs incurred by researchers in undertaking their analyses – whether these costs relate to compute consumed, or the access, ingress, or storage of additional datasets. UK Biobank’s working model would be for such charges to be billed directly to the researcher (or the research organisation they represent) without involvement from UK Biobank, and the platform provider should have a direct contractual relationship with them to enable this.

* Could you have the ability to manage the charging and billing of researcher use in the way envisaged?
* How could you address the need to charge back costs incurred by individual researchers?
* How could you ensure that the amounts charged to researchers remain competitive, transparent and fair?

**6. Commercial Model**

UK Biobank is keenly aware of the considerable interest the global research community has in the data it makes available, as evidenced by the growing numbers of registered researchers (~12,500), active research projects (~1,500), and published papers (~1,000). Given the visibility and significance of the resource, UK Biobank would be interested in commercial models that enable it to reduce the cost to itself of core platform provision, provided these did not significantly increase the costs incurred by researchers wishing to use such a platform service.

* Are there alternative commercial models for platform provision that UK Biobank could consider?
* Does the market have experience of working with any such models?

**7. Timeline and Feasibility**

UK Biobank is seeking to procure the platform during Q4 2019 with contract award in early 2020. The initial platform should be available in late Q1 2020 for beta testing, and in Q2 2020 for production use. The initial offering may be limited both in terms of scope of data held, types of analyses possible, and number of research groups to whom access is provided; but as a minimum, the platform must hold structured and WGS data, and must be suitable for genomic studies combined with these structured data. UK Biobank is considering a contract for the platform implementation on an initial fixed term basis with a rolling renewal.

* Could you fully implement your platform to meet all the requirements described in the timescale envisaged?
* If you do not consider that full implementation is possible in the timescale envisaged, what elements could be lacking initially and what could be the timescale for incorporating these?
* What key risks do you consider may be relevant?
* What dependencies do you consider may be relevant?
* What third party dependencies do you consider may be relevant?
* What would you consider may be a commercially appropriate structure for the term / renewability of contractual arrangements?

**8. Enabling Technologies**

UK Biobank recognises that there are a wide variety of technologies from which the platform might benefit (including but not limited to GPU hardware to accelerate compute, genomic-specific data compression to reduce storage requirements, and specialised database technologies to facilitate analyses). UK Biobank recognises that companies offering such niche technology solutions may not be able to offer the full range of capabilities the platform must provide, but nonetheless wishes to understand what might be available.

* Are there specific technologies UK Biobank could consider as part of its platform provision?
* How could UK Biobank and any 3rd party provider look to exploit these technologies?