UK Biobank Limited

Procurement Name:
Managed Informatics Platform for Research Access to Data

Procurement Reference Number:
UKBB009

Procurement Procedure:
Open

Invitation to Tender (ITT)

Specification

Contents

1 INTRODUCTION

1.1 Over the next five years, based on current scientific programmes, the UK Biobank resource will grow to approximately 15 Petabytes of data. UK Biobank researchers have previously downloaded data for analysis within their own environment. The increasing scale of the data requires a new approach to data storage and access that allows researchers to bring their analyses to the data through implementation of a managed informatics platform for data access and analysis (the 'Platform').

1.2 This Specification sets out UK Biobank's requirements for the initial provision of such a Platform, including:

- technical requirements for Platform functionality;
- implementation approach and testing; and
- service management and support.

1.3 UK Biobank is seeking a supplier (the 'Service Provider') that can provide such a Platform as a managed service with responsibility for all aspects of implementation and ongoing operational management.

2 GLOSSARY

This section provides a definition for certain terms used throughout the remainder of the Specification, some of which are specific to UK Biobank's operational context:

- **Access Application**: a request by a bona fide researcher to access UK Biobank Data or samples for research in the good of public health (and also denoting the ongoing project that arises from the approved request).
- **API:** Application Program Interface, specifying how software components shall interact using agreed routines and protocols.
- **Collaborators**: registered UK Biobank researchers working collaboratively with a Principal Investigator on an Access Application project.
- **FPGA**: Field-Programmable Gate Array (a specialised type of compute node).
- **GDPR** - General Data Protection Regulation 2016/679 is a regulation in EU law on data protection and privacy in the European Union and the European Economic Area.
- **GPU**: Graphical Processor Unit (a specialised type of compute node).
- **Healthcare record data**: codified data held by UK Biobank in pseudonymised form obtained through linkage to NHS and registry records.
- **Managed Service:** a service which comprises the necessary infrastructure, software, support and ongoing maintenance, with the necessary support personnel to deliver an operational service with appropriate monitoring and controls.
- **Pseudonymised Participant Identifier (EID)**: the participant identifier specific to a single approved Access application (referred to as the Encoded Identifier, or EID).
- **Pipeline**: a sequence of tools, with the output of one tool being automatically fed as the input of the next tool.
- **Platform**: a set of technologies on which applications/analyses can be developed.

- **Research institution**: the legal entity that has entered into a Material Transfer Agreement with UK Biobank in the context of an Access Application and to which Users are organisationally affiliated.
- **Service Provider**: a supplier who can deliver and operate a Platform, the storage and compute necessary to exploit it, and ancillary services to manage its use by others.
- **Terms and Conditions**: the terms and conditions contained in the agreement to which the Service Provider will supply the Managed Services and the Platform.
- **Tool**: a software program that processes a given input or set of inputs to produce a given output or set of outputs.
- **UK Biobank Data**: the pseudonymised data held within the UK Biobank resource on each of its 500,000 participants comprising, but not limited to, self-report data, linked healthcare record data, physical measures, biological assay data, imaging data, activity data, and derived variables.
- **User**: a UK Biobank registered researcher that has been given permission (by UK Biobank) to use the Platform.
- **Web Portal:** an internet facing website that provides access to information and services.
- **WGS Main Phase**: a project being run as an external Access Application led by a Consortium of 4 industry parties (with additional funding from government and charity). The project will sequence 450,000 genomes (in addition to those 50,000 genomes sequenced by UK Biobank within the WGS Vanguard project) with an initial tranche of 125,000 genomes by May 2020 and completion by early 2022.
- **Workflow**: an orchestration (or its codified description) of all the resources needed to execute a pipeline.
- **Workspace**: a project area that provides controlled access to data and allows authorised Users to work with those data.

3 OVERVIEW

3.1 This section provides an overview of UK Biobank's technical specification; each element is described in more detail in subsequent sections of this document.

3.2. UK Biobank is seeking a Service Provider that can deliver a Platform comprising of:

- **Access and Analysis Services** (see Section 4.1) – data access and analysis software to allow researchers to analyse UK Biobank Data in-situ.
- **Storage Services** (see Section 4.2) – to host and provide access to UK Biobank Data, researcher derived data within project workspaces, and additional datasets that researchers may upload from elsewhere.
- **Compute Services** (see Section 4.3) – to meet the needs of researchers who will perform a range of analyses, including complex computations requiring many compute cores, using general purpose, compute-intensive, memory-intensive and/or specialised compute instances (such as GPU or FPGA nodes).
- **Platform Support Services** (see Section 4.4) – to manage the Platform on behalf of UK Biobank, ensuring secure and auditable access to UK Biobank

Data, and providing other supporting services including charging and billing, and helpdesk as part of a managed service.

3.3. The following diagram outlines UK Biobank's considerations as to the major functional components and subcomponents of the Platform (see Figure 1). The diagram provides a framework to describe the technical requirements (as outlined in Section 4) and is not intended to be prescriptive of the architecture of any proposed solution.

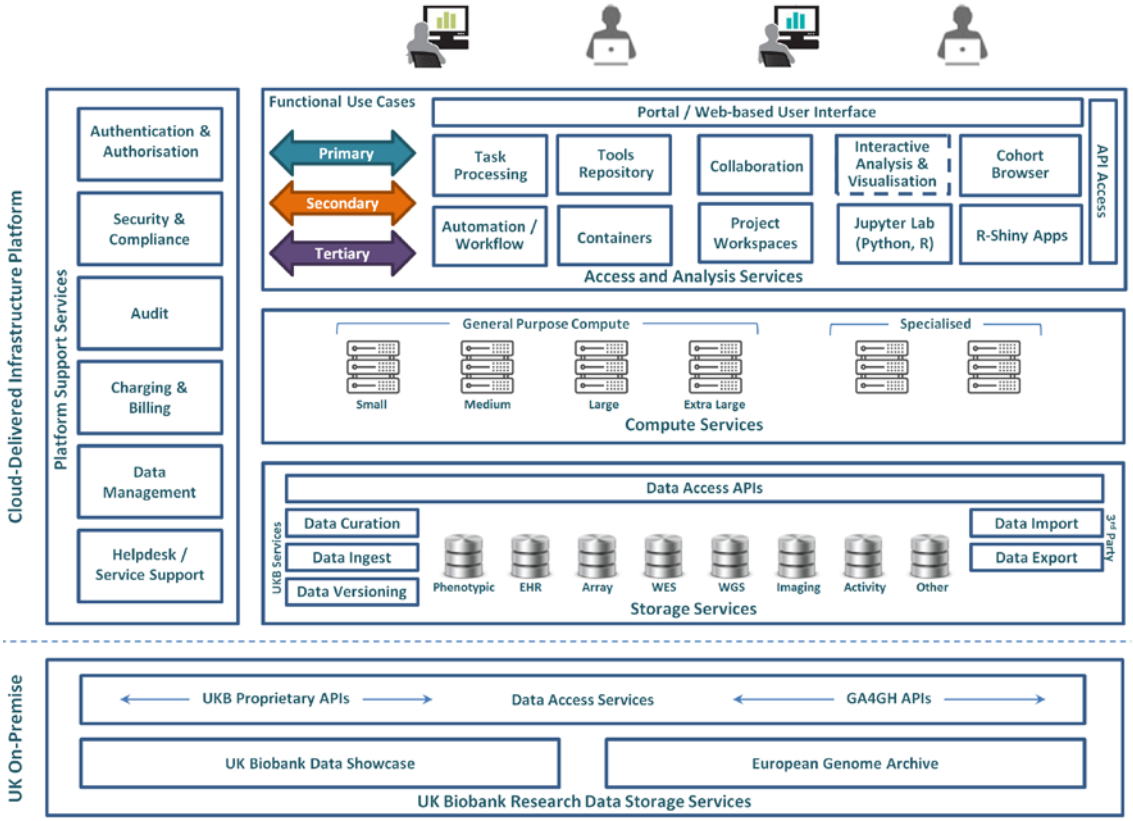

Figure 1. Platform functionality and existing UK Biobank data storage services

3.4. An initial version of the Platform is required as soon as possible after contract signature and, as such, UK Biobank is seeking customisation and/or configuration of an existing service offering, with the Service Provider responsible for all aspects of implementation as set out in Section 5.

3.5. Within the contracted term, there will be 2 discrete delivery phases, each providing an increase in the number of users able to access the Platform:

- **Phase 1 (Production Release Go-live – Limited availability to industry parties)**
  - **Timeline**: required as soon as possible after contract signature;
  - **Usage**: WGS Main Phase Consortium industry parties only.
- **Phase 2 (Production Release Go-live - General availability to all Users)**
  - **Timeline**: required 2Q 2021;
  - **Usage**: All UK Biobank registered researchers.

For further details of timescales, phasing and milestones see Section 5.2.

3.6. Subsequent to the initial implementation, there shall be iterative Platform releases on a periodic basis (typically bi-annually) to incorporate updates of UK Biobank Data (such as more participant imaging datasets, or new data types arising from subsequent UK Biobank enhancement projects).

3.7. Given the timescales to be met, UK Biobank anticipates the Platform will be a UK Biobank-specific instance of an existing and more widely available platform product, and that the Service Provider will continue to develop its offering over time. As such, it is required that any such development becomes routinely available as part of the UK Biobank-specific instance. UK Biobank requires that it is able to provide input to the product roadmap alongside the Service Provider's other clients.

3.8. The Service Provider will be responsible for formal testing (including performance testing) as part of each implementation phase (and for subsequent Platform releases) as detailed in Section 5 and in accordance with Schedule 6 of the Terms and Conditions.

3.9. The Service Provider shall deliver the Platform as a managed service on behalf of UK Biobank and will be the first point of contact for dealing with Platform-related issues. The Service Provider will be responsible for service reporting, delivery of helpdesk, management of problems/issues, on-boarding of new users and provision of education and support materials as set out in Section 6.

3.10. The Platform service will be used by approved Users who are located all around the world, and shall therefore provide service availability to meet the needs of researchers in different time zones (note: only English locale language support is required). The Service Provider shall be able to ensure that they can meet the Performance Levels in accordance with Schedule 3 of the Terms and Conditions.

3.11. UK Biobank is seeking to contract with a Service Provider who will be responsible for the end-to-end Platform service. At this time, UK Biobank is not seeking to contract directly with sub-contractors who may provide component parts of the service, such as the provision of compute and storage services, and rather expects the Service Provider to manage the contractual and operational dependencies on any subcontractors. However, UK Biobank will retain the right to novate such subcontractor relationships at the end of the contracted term in accordance with the provisions of the Terms and Conditions.

3.12. In the longer term, an ecosystem of multiple Service Providers and infrastructure providers would allow researchers the flexibility 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. UK Biobank may therefore run further tenders for Platform functionality in the future (which may be before or after the end of the contracted term).

4 TECHNICAL REQUIREMENTS

This section outlines UK Biobank's technical requirements for the Platform.

Appendix B to this Specification sets out those requirements that need to be delivered as part of Phase 1 versus Phase 2 of the Platform implementation.

4.1 Access and Analysis Services

4.1.1 Web Portal and API Access

Users shall be able to access the data and analysis services the Platform provides both through a Web portal providing direct interactive access to Users, and through a Platform API providing programmatic access, subject in both cases to appropriate User authentication and authorisation.

The Web portal shall be accessible through widely supported Web browsers. The Web portal branding shall be configurable to reflect UK Biobank iconography.

The Web portal shall allow Users to register as Users and, having done so, allow them to create workspaces (see Section 4.1.2) within which they can access and analyse data. The Platform shall integrate with UK Biobank's OAuth Web service to determine whether an individual is an approved researcher and to verify that they are part of a specific UK Biobank Access Application (see Appendix A.3).

The Platform API shall support invocation of Platform services for data access and analysis (including execution of analyses, such as workflows or pipelines, that Users define, store, and run within the Platform environment) via command line tools (and for which the Service Provider may provide an SDK offering such tools).

4.1.2 Workspaces

Each UK Biobank approved Access Application shall have read-only access to a subset of the UK Biobank Data that has been keyed using a set of UK Biobank-supplied pseudonymised identifiers (EIDs) specific to the Access Application (see Appendix A.1 for more detailed requirements and examples of how this process shall work). The Platform shall integrate with UK Biobank Web services to determine the correct subset of UK Biobank Data and specific EIDs for each Access Application.

UK Biobank Data shall be keyed using the set of EIDs specific to the User's approved Access Application, and be accessed for review, analysis, or download (if permitted) within the context of a virtual project area or workspace that:

- controls access to UK Biobank Data or other data;
- allows these data to be analysed by Platform-provided or other tools;
- allows these data, analysis code and/or results to be appropriately shared with Collaborators.

The Platform shall regard the UK Biobank Data it holds as a read-only master dataset and, when these data are made available through a workspace, it shall be done via

linking to the Platform's master copy of the data rather than by creating an additional physical copy within in the User's own workspace.

The Platform shall be able to manage permissions at both the Access Application level, and at the more granular data field-level that determines which UK Biobank Data the User is allowed to access.

The Platform shall:

- Allow Users to create multiple workspaces;
- Allow Users to associate an individual workspace with one (and only one) UK Biobank Access Application;
- Confirm that the User is an approved researcher on the Access Application (otherwise access shall be denied) both at the time of initial association and on subsequent access; and
- Only allow UK Biobank Data to be copied (or shared) between workspaces associated with the same Access Application.

UK Biobank will provide an OAuth Web service that allows the Platform to confirm User permissions before UK Biobank Data are initially linked to the workspace and for each subsequent access session through the Platform (whether via the Web portal or API).

The creator of a workspace shall be able to invite other Users to join it; the Platform shall ensure that such invitees are Collaborators on the same Access Application.

The creator of the workspace shall be able to add charging and billing details such that any chargeable Platform resources consumed are tracked and appropriately charged back to the User (see Section 4.4.4).

The workspace permissions shall allow Users to define those other Users who can make changes within each workspace (such as inviting additional Users to join the workspace, or changing charging and billing information).

4.1.3 Interactive Analysis (including Visualisation)

The Platform shall provide the ability for Users to interactively explore, analyse and visualise the data, and support a variety of user interfaces based upon standard bioinformatics research tools, including (but not limited to):

- a **Cohort Browser** - to support data browsing and sub-cohort definition based upon phenotypic, genotypic or other participant characteristics:
    - it shall be possible to define multiple sub-cohorts based on user-defined phenotypes, and to save and share those definitions.
    - it shall be possible to use any attribute or data element as a sub-cohort selection criterion. As each is selected, the number of participants for whom data are available shall be indicated, and (where relevant) an indication of the data distribution shown. The Cohort Browser shall only access those data that the User is authorised to view for that Access Application.

- an example of a simple sub-cohort definition might be for participants with a specific ICD-10 code (such as for vascular dementia, F02*) recorded in healthcare record data and for whom brain MRI imaging data are also available.

- a **Notebook** capability (such as that provided by Jupyter Notebooks[1]) for Users who may wish to develop their own code for data analysis:
  - support shall be included for common programming languages used for data analysis, including Python and R;
  - support shall be included for Notebooks that require access to distributed processing (e.g. Apache Spark) for the running of specialised tools, such as HAIL[2].

In addition to an interactive Cohort Browser and Notebook capabilities, the Platform shall provide support for the development, upload, and deployment of other interactive applications, for example, graphical tools appropriate for viewing genetic data and results for visual validation of genetic variants such as Broad's IGV[3] or UCSC's Genome Browser[4], or outputs from association studies such as a GWAS browser for reviewing GWAS and/or PheWAS Manhattan plots and interrogating specific regions using LocusZoom.

4.1.4 Batch Analysis (including Tools and Pipelines)

The Platform shall support a range of analytical use cases, such as statistical analyses, batch-level processing of raw genetic data to generate high quality variant call-sets, conducting GWAS or PheWAS association studies, and other analyses that researchers may wish to undertake including:

- Phenotype analyses – e.g. disease prevalence analyses and definition of computable phenotypes extracted and cleaned from health record data and/or image-derived phenotypes;
- Genotype analyses – e.g. extraction of genotypes/variants from raw read data, and annotation of variants / gene / pathway function and/or impact;
- Association analyses – performing individual phenotype GWAS (including composite genotypes like LoF burden) and individual genotype PheWAS (including custom/composite phenotype definitions, and meta-analyses across datasets); and
- Translational analyses – defining and capturing detailed phenotype definitions based on inclusion/exclusion criteria.

Commonly required tools and pipelines to support the above shall be predefined as part of the Platform, but it shall also be possible for Users to define their own

[1] Jupyter Notebooks - an open-source web application that allows creation and sharing of documents that contain live code, equations, visualisations and narrative text - https://jupyter.org/
[2] HAIL – a python based library for analysing genomic data - https://hail.is/
[3] Integrative Genomics Viewer (IGV) - https://software.broadinstitute.org/software/igv/
[4] UCSC's Genome Browser, available in format such as Genome in a Box - https://genome-store.ucsc.edu/

pipelines which incorporate both Platform native tools and their own uploaded tools.

The tools and pipelines available as part of the Platform shall include, but not be limited to, those that support:

- standard statistical analyses such as Stata, SPSS, SAS, R (subject to appropriate licensing);
- common processing of genomic data (such as GATK);
- WGS and WES variant calling from sequence data;
- manipulation of variant call file outputs (such as bcftools or samtools);
- analysis of alignment variation;
- association studies, such as GWAS and PheWAS analyses;
- analysis of population structure and relatedness;
- polygenic risk scoring;
- image processing, for derivation of image-derived phenotypes; and
- machine learning.

Users shall be able to run analysis processes by invoking them directly through the user interface, by including them in analytic pipelines, or by invoking them programmatically using the Platform API.

4.1.5 Tool Repository

The tools and pipelines available within the Platform shall be described in, and accessible through, a searchable repository which shall be consistent with the patterns set out in the GA4GH Tool Registry Service API[5].

Tools and pipelines that are part of the core Platform offering (rather than being uploaded by Users) shall be optimised to make best use of the Platform services and underlying cloud infrastructure.

Tools and pipelines shall be version controlled to support change control and reproducibility. When new versions of tools and pipelines are released as part of the Platform, access to previous versions shall be easily (and identifiably) available to allow researchers to assess changes between versions and impact on analyses they have conducted previously. Version information for the tools and pipelines used in a specific analysis shall be logged to support reproducibility.

The Platform shall be extensible and provide capabilities for allowing Users to use additional tools and pipelines that are not part of the standard Platform offering. It shall be possible for Users to incorporate their own tools and pipelines using container technologies such as Docker or Singularity. The Platform shall limit any damage caused by running malicious code solely to the workspaces associated with the end User running it.

[5] The Tool Registry Service (TRS) API is one of a series of technical standards from the GA4GH Cloud Work Stream that allow researchers to bring algorithms to datasets in disparate cloud environments - https://ga4gh.github.io/tool-registry-service-schemas/

Tools and pipelines shall be shareable with other researchers associated with the same UK Biobank Access Application, with other members of the researcher's institution, with their collaborators, or be made public.

It shall also be possible for tools developed by others elsewhere (subject to appropriate licensing) to be accessed from and used within the Platform, such as tools hosted within other repositories such as Dockstore[6] or BioContainers[7].

The Service Provider shall maintain a process for the prioritisation and incorporation of additional tools based on User demand and priorities agreed with UK Biobank, to allow new tools to be added to the Platform over time.

4.1.6 Automation (including Workflow and Process Execution)

The Platform shall support one or more standard workflow languages (such as CWL, WDL or Nextflow[8]) for pipeline definition and automation.

The Platform shall include a process execution engine to run defined pipelines and workflows, and shall be able to scale workflows vertically and horizontally, and where compute instances can be defined as part of workflow definitions.

This process execution engine shall be consistent with the patterns set out in the GA4GH Workflow Execution Service API.[9]

4.1.7 Usability

There are a number of distinct user groups whose needs the Platform shall be able to meet. These groups vary in their level of technical ability, their area of research focus, the data they wish to explore, and the analyses they wish to undertake. They include (for the purposes of illustration only):

- Pharma researchers who may want to use their own pipelines against data from multiple sources;
- Experienced academics, who may want to work with data using their preferred software tools and programming languages; and
- Doctoral students and early career researchers, who may want to engage directly with the data through a visual platform.

The initial Phase 1 implementation shall meet the needs of the Pharma companies who comprise the WGS Main Phase Consortium, and who may be considered to have a level of technical familiarity with the tools and analyses the Platform supports.

---

[6] Dockstore is an online repository where users can share tools encapsulated in Docker and described with the Common Workflow Language (CWL) or Workflow Description Language (WDL) to enable scientists to share analytical tools in a way that makes them machine readable and runnable in a variety of environments - https://dockstore.org/

[7] BioContainers is a community-driven project that provides the infrastructure and guidelines to create, manage and distribute bioinformatics packages (e.g conda) and containers (e.g docker, singularity) - https://biocontainers.pro/

[8] Workflow Description Language (WDL), Common Workflow Language (CWL) and Nextflow: examples of workflow definition standards

[9] The Workflow Execution Service API is one of a series of technical standards from the GA4GH Cloud Work Stream that describes a standard programmatic way to run and manage workflows to let people run the same workflow using various execution platforms running on various clouds/environments - https://ga4gh.github.io/workflow-execution-service-schemas/

By Phase 2 of the implementation, the Platform shall be capable of meeting the needs of the full range of UK Biobank researchers, and be usable by the research community whatever their interest or level of technical expertise.

The user interface shall follow industry good practice in usability design and follow established usability principles:

- **Visibility**: everything to complete a task shall be available and apparent when and where needed;
- **Feedback**: users shall be kept informed of consequences of actions, events, and progress relevant to them and their tasks;
- **Structure**: layout shall be organised by meaning and use and shall make sense to Users in terms of their intentions;
- **Reuse**: purposeful consistency shall be maintained through reuse of internal and interface components and behaviours;
- **Tolerance**: the user interface shall support flexibility in interactions, allowing Users to complete tasks in a number of different ways and catering for potential User errors;
- **Simplicity**: important and frequent tasks shall be simple and straightforward.[10]

The user interface shall support both new and established (expert) users to accomplish tasks efficiently and with minimal interactions. Contextual online help shall be provided, with links to online documentation and tutorials.

The user interface shall be accessible, and support standards such as the Web Content Accessibility Guidelines (WCAG).

The Service Provider shall demonstrate that it considers usability as an underlying principle within its software development lifecycle, which shall include addressing usability requirements through techniques such as User Centred Design and undertaking usability testing as part of the test cycle.

4.2 Storage Services

4.2.1 Data Storage

The Platform shall provide hosting for, and controlled access to, a copy of all UK Biobank Data that have been made available to approved Access Applications, including:

- structured (tabular) data comprising phenotypic data (i.e. field-level data, such as health-related data collected at baseline) and linked healthcare record data (i.e. codified data extracted from longitudinal health records and registries). These structured data are currently held in one or more relational databases and are less than 1 TB in size; and
- bulk (BLOB[11]) data such as MRI images and genetic sequences.

[10] Lockwood and Constantine: Usability by Inspection (2003)

The estimated data volume projections for the forthcoming period based on committed scientific programmes are set out in Table 1 below.

	Up to 2019	2020				2021				2022				2023			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Cumulative Compressed (in TB)	**383**	**2849**	**4020**	**5791**	**7812**	**9333**	**10854**	**12380**	**13906**	**14432**	**14458**	**14483**	**14508**	**14533**	**14558**	**14583**	**14608**
Structured (baseline, healthcare)	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Whole Genome Sequencing	0	2000	3000	4500	6000	7500	9000	10500	12000	12500	12500	12500	12500	12500	12500	12500	12500
Exome Sequencing	150	600	750	1000	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500
Genotyping/Imputation	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
Imaging	185	200	220	240	260	280	300	325	350	375	400	425	450	475	500	525	550
Cardiac Monitoring	2	3	4	5	6	7	8	9	10	11	12	12	12	12	12	12	12
Activity Monitoring	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
Recruitment	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20

2020				2021				2022				2023			
Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
2849	4020	5791	7812	9333	10854	12380	13906	14432	14458	14483	14508	14533	14558	14583	14608

Table 1. Data storage volumes (in TB) by data type.

Note: the volumes for Whole Genome and Whole Exome Sequencing comprise variant (VCF) and sequence (CRAM) data; variant data are ~25% of the total volume.

The data storage system shall maintain granular access permissions to cater for scenarios where some Access Applications only have permission to use certain subsets of UK Biobank Data.

Further details relating to the specification of the UK Biobank Data held are set out in Appendix A.2.

The Platform shall provide tools to monitor and report on data access and usage.

4.2.2 Data Management

The Platform shall support UK Biobank's strict pseudonymisation approach from the outset, to ensure that researchers can only interact with UK Biobank Data in the context of their approved Access Application, as set out further in Appendix A.1.

The Service Provider will be responsible for ingesting new data into the Platform to be made available to Users at regular intervals, typically bi-annually, and shall be able to ingest steady streams of data, such as assay data being generated as part of the WGS Main Phase project. Future data releases may include new data types (such as proteomic and metabolomic data), the inclusion of additional imaging data as further participants are scanned, or refreshes of linked electronic health record data.

UK Biobank will retain responsibility for communicating details of new and expanded datasets within each release. The Service Provider will be responsible for also making these details available to Users through tools and/or documentation provided by the Helpdesk (or within the Platform).

The Platform shall support removal of participant-related data following notification from UK Biobank of any participant withdrawals (though the number of withdrawal requests remains very low). Such removal shall encompass deletion from the Platform copy of the UK Biobank Data those individual records and files related to

[11] Binary Large OBject (BLOB) is a collection of binary data stored as a single entity in a database management system, typically images, audio or other multimedia objects

the participant(s) who have withdrawn, and their removal from shared files such as multi-sample PLINK files or pVCFs (either through direct manipulation of the files or their replacement with updated copies obtained from UK Biobank).

Details of participants who have withdrawn will be provided to the Service Provider on a regular basis (typically quarterly), and data relating to these withdrawn participants shall be deleted upon receipt of this notification (and the Service Provider will be responsible for providing appropriate communication to Users to accommodate such datasets changing as part of their analyses).

In the course of their research, the Users generate results data which underpins their research findings and which they are obliged to return to UK Biobank. Where a User generates derived variables that are to be returned, the Platform shall allow these data (together with accompanying metadata and documentation) to be copied to a UK Biobank staging area (or equivalent) within the Platform where they may be reviewed.

4.2.3 Data import and export services

Users shall be able to upload other data directly related to their Access Application , and shall be able to undertake combined analyses that include their own or other publicly available datasets for which such linkage is technically feasible. Such upload and linkage shall be accomplished using industry standard tools (e.g. FTP, SFTP) and open standards (e.g. GA4GH APIs). Additional proprietary tools shall also be made available if they offer particular advantage, for example in terms of data transfer speed. The requirement for using industry standard tools (e.g. FTP, SFTP) also extends to data download and export.

All such access shall be subject to appropriate levels of User authentication and access control.

Users shall be able to download and export UK Biobank Data (subject to the restrictions noted below) from the Platform for further analysis using their own infrastructure.

UK Biobank may wish to restrict the ability to download or export certain UK Biobank Data from the Platform and such restrictions shall be supported; this might involve, for example, review of any request to download data in excess of 100GB.

Users shall be able to download the results of their analyses to their own servers from the Platform

4.3 Compute Services

UK Biobank considers the most appropriate infrastructure on which to deliver the proposed Platform is widely available public cloud infrastructure, which offers the necessary flexibility and elasticity to accommodate multiple large-scale requests to access and analyse the data.

UK Biobank requires the Service Provider to be able to make available a range of compute instances that can be selected by the User based on their analysis requirements. Such analysis may be relatively simple, may require significant computational capability, may require significant amounts of memory, and/or may benefit from specialised computing capabilities.

The instance types to be available (both Linux and Windows) shall include:

- general purpose compute instances;
- compute-intensive instances;
- memory-intensive instances; and
- specialised compute instances (such as GPU or FPGA nodes).

The User shall be able to select from a standard list of available configurations in each of the categories set out above (with recommended configurations suggested for typical analyses or use cases), and shall be able to customise their own configuration (for allocation of numbers of core, memory and/or storage).

Service Providers shall identify any configuration limits that may apply (such as maximum memory or core allocation), and any particular types of analysis whose requirements are anticipated to exceed the resources that can be made available to a researcher.

UK Biobank will be reliant on the Service Provider's experience to predict likely capacity requirements and variability in demand (particularly given the complexity and non-trivial nature of genomic analyses). For example, to remap and call 1,000 WGS samples would likely require the order of 1,000,000 core hours; to joint call 50,000 WES for a single chromosome using GATK can consume nearly 2,000 core hours[12].

The Platform shall be able to support such analyses being run by multiple Users concurrently. In addition, there will likely be transient cases of very high levels of use (e.g. when new data are released) where queuing mechanisms will be required to spread workload demands.

Platform capacity and usage patterns shall be closely monitored, particularly during the initial implementation and subsequent releases.

[12] GLnexus: joint variant calling for large cohort sequencing - Michael F. Lin, Ohad Rodeh, John Penn, Xiaodong Bai, Jeffrey G. Reid, Olga Krasheninina, William J. Salerno – bioRxiv 343970; doi: https://doi.org/10.1101/343970

4.4 Platform Support Services

4.4.1 Authentication and Authorisation

The Platform shall include authentication and authorisation controls so that only Users are able to access the data, and shall build upon the capabilities of the underlying infrastructure service to ensure data integrity and availability.

Users shall register with the Platform to establish a Platform account. The Platform shall confirm that potential users of the Platform are UK Biobank approved researchers before allowing them to register (and UK Biobank will provide an OAuth Web service for this purpose).

For the Phase 1 implementation, UK Biobank's preference is for the OAuth service to be used from the outset. However, given the small number of research institutions and limited number of Users, UK Biobank may accept an alternative approach to de-risk and accelerate the Phase 1 implementation where these account credentials are specific to (and held within) the Platform. In any event, UK Biobank would require full OAuth functionality to be in place by Phase 2.

The APIs to be provided by UK Biobank are described further in Appendix A.3.

4.4.2 Security and Compliance

The UK Biobank Data to be stored within the Platform are pseudonymised data (the data is de-identified when it is provided to Users: the re-identification keys are never provided to Users). As such, UK Biobank considers that it is not reasonably possible for Users of the Platform to identify a UK Biobank participant from the provided UK Biobank Data even taken in conjunction with reasonably available public data.

Nevertheless, UK Biobank considers that it still has a responsibility to ensure that the integrity surrounding storage and access to the data is maintained. To this end, UK Biobank requires the Platform to process and store the data as if it were personal data (in other words to the highest standards of professional probity and security).

Further, if the UK Biobank Data provided (or component parts thereof) are in due course considered to be personal data (including special category personal data) then a suitable set of standardised compliant controller / processor clauses (set out in the Terms and Conditions) will apply to the Service Provider (on UK Biobank's notification). UK Biobank requires the Service Provider to act on UK Biobank's instructions and to deliver the Platform in a manner that remains compliant with the GDPR at all times.

Further, there are additional clauses – which are compliant with the requirements of the Information Commissioner's Office [https://ico.org.uk/] and the GDPR – which provide an appropriate legal basis for a transfer of the data outside of the European Economic Area. In any event, UK Biobank may require the Service Provider to promptly disable access to UK Biobank Data from a particular storage node(s) and/or only store UK Biobank Data in particular nodes.

Furthermore, the extent of the Service Provider's obligations will extend beyond the processing of UK Biobank Data and may also include the processing of personal data relating to Users of the Platform (including for the purposes of managing access credentials, contact details, billing information and usage data for audit purposes).

Specifically, UK Biobank will require certain confirmations from the Service Provider, including:

- The Service Provider shall notify UK Biobank without undue delay in the event of any security or data breach being identified;
- The Service Provider shall maintain regular penetration testing and software / infrastructure vulnerability assessments (and where appropriately summarised reports from such assessments can be made available upon request from UK Biobank, and where such requests will not be unreasonably withheld);
- The Service Provider shall have an appointed Data Protection Officer who would be the first point of contact for UK Biobank to address data protection-related concerns; and
- The Service Provider shall be able to demonstrate that it maintains appropriate security controls (encompassing the scope of the Platform including its underlying infrastructure) through certification to ISO27001 (or an equivalent standard, such as FedRAMP assessment) and that the following requirements will be met:
    - only authorised users shall be able to access the Platform and/or the data available within the Platform
    - data shall be encrypted at rest and in transit (when transitioning outwith the Platform) using AES-256 (or equivalent) encryption
    - there shall be selective access controls that allow, for example, Users to be restricted to specific datasets (or subset thereof) depending upon the context (or workspace) that they are working within
    - there shall be controls in place to securely remove data when it is no longer needed (for example, at the end of the service contract); and
    - the Platform shall limit any damage caused by running malicious code solely to the workspaces associated with the User running it.

4.4.3 Audit

An audit trail shall be maintained of Platform activity, covering data access and download. Audit data shall be retained for a minimum of 6 years for regulatory compliance, and a copy made available to UK Biobank for archiving prior to deletion by the Service Provider.

Audit logs shall be reviewed regularly by the Service Provider to identify anomalous activity such as excessive failed logins, and any such suspicious activity shall be reported to UK Biobank.

Audit logs shall be reviewed as necessary to investigate suspected instances of unauthorised data access or data leak, either at the instigation of the Service Provider or on request by UK Biobank.

4.4.4 Charging & Billing

UK Biobank requires that the costs associated with the storage of the Platform copy of UK Biobank Data and any ongoing Platform service charge(s) associated with the operational running of the core Platform will be borne by UK Biobank. UK Biobank expects to be invoiced for these costs monthly in arrears.

All other costs associated with compute utilisation, storage of additional data (whether imported from external sources or derived from analyses undertaken on the Platform), and any data ingress/egress shall be borne by individual Users.

The Platform shall allow limits to be set on the charges that can accrue at the workspace, User, or research institution level. Such limits shall be set by specifying a maximum set of resources (cores, memory) and/or a maximum cost per month.

Resource consumption shall be logged by User/workspace, and the Platform shall allow Users to be linked to organisational entities (such as a research institution), and billing across an entity to be monitored and controlled.

It shall be evident to a User what resources will be used as part of specific analyses, with guidance on the costs that may be incurred provided in advance and/or provide the ability for Users to cap costs by stopping an analysis job if a threshold is exceeded. If there are multiple ways of servicing a request (for example, trading time taken to complete against cost through the use of on-demand versus cheaper rate compute instances), these shall be indicated and the User allowed to select the most appropriate approach.

Similarly, where data are to be uploaded or downloaded by the User, an indication of the costs associated with their ingress, storage, or egress shall be provided.

Users shall register with the Platform to establish a Platform account and provide billing details. They shall not necessarily be required to enter billing details at the point of account creation, but shall do so before consuming any chargeable Platform resources. The creator of a workspace shall be responsible for all charges associated with it, and shall be able to review resource consumption by each User accessing the workspace.

Users shall be invoiced for charges they have incurred directly, and the Platform shall support a variety of payment methods by which Users may settle their account, including credit card billing and institutional purchase orders.

UK Biobank will continue to manage the Access application process with Users and the access fees associated with this: this remains a separate process to the Service Provider charging the User for their utilisation of Platform resources.

5 IMPLEMENTATION

5.1 Governance and Project Management

5.1.1 The Service Provider shall identify an Executive Sponsor for the project who will oversee and be accountable for project execution, and who will act as an escalation point for issues that cannot be resolved at the working project level. The Executive Sponsor shall hold regular reviews with their UK Biobank equivalent.

5.1.2 An Executive Steering Group involving members of the UK Biobank Executive team (and others as UK Biobank may deem to invite) and appropriate members of the Service Provider shall meet monthly during the initial period and quarterly thereafter to oversee satisfactory progress, timely implementation, operational performance and future development plans (considering and prioritising feedback from the UK Biobank research community).

5.1.3 The Service Provider shall assign a named Project Manager with relevant qualifications and experience of delivering similar projects, who will act as the principal point of contact for management of the implementation project: including delivery of all project phases and any subsequent Platform releases, ingestion of UK Biobank Data, and coordination of public data releases.

5.1.4 The Service Provider Project Manager will be responsible for:

- Project management tasks, including:
  - project setup activities;
  - development of project documentation; and
  - oversight of the Platform implementation;
- Development of Standard Operating Procedures for service delivery, including:
  - data management: data ingestion, participant withdrawals, data releases;
  - helpdesk: support and exception handling; and
  - communication: platform adoption, UK Biobank staff and User engagement.

5.1.5 The Service Provider Project Manager shall develop (for review and approval by UK Biobank) a detailed project plan for each phase of the Platform implementation, including setup of service management and ingestion of UK Biobank Data, and such subsequent development/release phases as may be needed to address UK Biobank's functional requirements, in accordance with Schedule 5 of the Terms and Conditions.

5.1.6 The Service Provider shall have a defined approach to project management, either via the Service Provider's own project management framework (evidenced with reference to its successful use on projects of a similar scale and complexity) or derived from an industry standard project management framework (e.g. Prince2).

5.1.7 Any dependencies on UK Biobank infrastructure, systems, or staff that need to be met to facilitate implementation and/or use of the service shall be identified.

5.2 Timescales, Phasing, and Milestones

5.2.1 There are two distinct phases to the project, each split into two sub-phases with the following timescales and milestones:

Milestone	**Deliverables** (*bulleted list showing all Deliverables (and associated tasks) required for each Milestone*	**Duration** (*Working Days*)	**Milestone Date**
1	Contract signature		Effective Date
2	Project Mobilisation: • Service Provider responsible for Project Definition Workshop to inform completion of Detailed Implementation Plan	5	One week after Effective Date
3	Delivery of Detailed Implementation Plan as set out in Section 3 of Schedule 5 of the Terms and Conditions	10	Two weeks after Effective Date
4 (Phase 1 ATP)	Phase 1 Beta Release Go-live: • Initial deployment of the UK Biobank Platform • Delivery of Phase 1 functionality as set out in Appendix B of the Specification • Ingestion and availability of the UK Biobank Data as set out in Appendix C of the Specification • On-boarding and access to the Platform for UK Biobank staff and WGS Main phase industry parties		As soon as possible after Effective Date
5 (Phase 1 CPP)	Phase 1 Production Release Go-live: • Phase 1 Production release of the UK Biobank Platform • Delivery of Phase 1 functionality as set out in Appendix B of the Specification • Ingestion and availability of the UK Biobank Data as set out in Appendix C of the Specification • Successful completion of testing as set out in Section 5.3 of the Specification and Schedule 6 of the Terms and Conditions		As soon as possible after Effective Date, and in any event during Q3 2020

Milestone	Deliverables (*bulleted list showing all Deliverables (and associated tasks) required for each Milestone*	Duration (*Working Days*)	Milestone Date
	• On-boarding and access to the Platform for UK Biobank staff and WGS Main phase industry parties		
6 (Phase 2 ATP)	Phase 2 Pilot Release Go-live: • Delivery of the functionality required for Phase 2 as set out in Appendix B of the Specification • Ingestion and availability of the UK Biobank Data as set out in Appendix C of the Specification • On-boarding and access to the Platform for additional approved researchers as identified by UK Biobank.		As soon as possible following Phase 1 Production Release
7 (Phase 2 CPP)	Phase 2 Production Release Go-live • Delivery of the functionality required for Phase 2 as set out in Appendix B of the Specification • Ingestion and availability of the UK Biobank Data as set out in Appendix C of the Specification • Successful completion of testing as set out in Section 5.3 of the Specification and Schedule 6 of the Terms and Conditions • General availability and access for all approved UK Biobank registered researchers (with on-boarding determined by User demand)		As soon as possible following Phase 2 Pilot, and in any event during Q2 2021

5.2.2 The Platform is required to be available as soon as possible after contract signature with a Phase 1 Production Platform for use by the industry parties as part of the WGS Main Phase Consortium. Timeliness of implementation is of critical importance to UK Biobank.

5.2.3 The Service Provider shall provide an outline implementation plan for each phase as part of their submission. The outline implementation plan shall include, as a minimum, the milestones set out in Schedule 5 of the Terms and Conditions and shall cover all activities from contract signature to go-live, including but not limited to activities associated with:

- project mobilisation;
- set up / configuration of the UK Biobank Platform;
- establish and test connectivity to UK Biobank API Web services (as required);
- development (as required) to meet Specification functional requirements;
- data ingestion for UK Biobank Data;
- testing of Platform functionality and data availability;
- Phase 1 beta release;
- training for UK Biobank staff (as may be required);
- on-boarding the WGS Main Phase parties;
- Phase 1 Production release.

5.2.4 As part of Project Mobilisation for Phase 1, the Service Provider shall be responsible for running a Project Definition Workshop with UK Biobank within one week of contract signature, and a detailed implementation plan developed and submitted to UK Biobank within two weeks of contract signature as set out in Schedule 5 of the Terms and Conditions.

5.2.5 For the Phase 2 Pilot, the Platform will be made available to a small number (<100) of UK Biobank registered researchers spread across a small number of research institutions (<10) in addition to the WGS Main Phase industry parties, to inform how well the Platform and service management scales with the number of Users (including on-boarding, self-enablement, service support and reporting) and gain feedback on Platform functionality to inform further enhancements. The pilot will run for a period of circa 3 to 4 months to gain feedback ahead of the Phase 2 Production go-live and general availability for all UK Biobank registered researchers. Feedback will be sought by UK Biobank from those researchers accessing the Platform during the Pilot, and following review fed back to the Service Provider for action (as appropriate).

5.2.6 As part of User adoption, the Service Provider shall support UK Biobank in the development of a User engagement plan to encourage the adoption of the Platform by the relevant members of the UK Biobank research community.

5.2.7 **Exit Management**. Prior to the end of the contract, the Service Provider shall create an inventory of the UK Biobank and User data and software held on the Platform, and work with UK Biobank to develop and execute an exit plan in accordance with Schedule 8 of the Terms and Conditions.

5.3 Testing

Service Provider Testing

5.3.1 The Service Provider shall be responsible for comprehensive testing (as expected within standard software and product development practices) as part of the service delivered to UK Biobank. The Service Provider shall ensure (and evidence) that the Platform is free from known critical defects on initial implementation and all subsequent releases, in accordance with Schedule 6 of the Terms and Conditions.

5.3.2 For each new release of the Platform, the Service Provider shall provide documentation of new functionality, issues resolved and known defects prior to release to UK Biobank.

5.3.3 The Service Provider shall follow a test approach aligned to industry good practice for software development, which shall include:

- creation of a test strategy, test plans and test cases and expected results;
- test execution, and maintenance of a test issue log; and
- production of a test exit report.

5.3.4 Testing shall confirm the base capability of the Platform to:

- support both interactive and API access;
- support workspaces, interactive and batch analysis, and automation;
- allow Users to upload their own code and data;
- allow analysis results to be saved, viewed, and downloaded;
- maintain an audit trail of activity; and
- allow multiple concurrent Users and maintain acceptable performance.

5.3.5 Testing shall also encompass functionality specific to UK Biobank:

- only UK Biobank approved researchers can register for Platform access;
- Users can only access UK Biobank Data for which they are approved;
- data are only accessed using the appropriate pseudonymisation scheme; and
- data may only be downloaded, if such download is permitted.

5.3.6 During testing, the Service Provider shall make available a copy of the test issue log and any other such test documentation as may be requested.

5.3.7 On completion of testing, the Servce Provider shall make available a test exit report which includes details of tests undertaken, tests successfully completed, defects identified and corrected, and any issues outstanding.

5.3.8 If, on review of the test exit report, UK Biobank deems any outstanding issue to be sufficiently critical to delay implementation, the Service Provider shall agree with UK Biobank a plan to correct the issue or provide a workaround.

UK Biobank Testing

5.3.9 In addition to reviewing Service Provider test results as outlined above, UK Biobank will undertake its own testing of each new Platform version (and each new data release) to confirm:

- basic operation of the Platform;
- adherence to authentication and authorisation requirements; and
- appropriate availability of UK Biobank Data.

5.3.10 The Service Provider shall provide a capability for UK Biobank to conduct such testing from UK Biobank premises in such a way that the existing production Platform service is not compromised.

5.3.11 The Service Provider shall review all defects reported by UK Biobank, and either provide a fix or agree with UK Biobank a plan for resolution (which might include agreement of a workaround and a commitment to fix in the next Platform release).

5.3.12 UK Biobank may wish to invite selected approved researchers to undertake usability testing of the Platform, and the Service Provider shall provide a capability for such testing to be undertaken from the researcher's own location in such a way that the existing production Platform service is not compromised.

6 SERVICE MANAGEMENT AND SUPPORT

6.1 Service Management

6.1.1 The Service Provider shall assign a named Service Manager who shall be responsible for operational delivery of the service and act as the Service Provider's principal point of contact for service management and support.

6.1.2 UK Biobank will identify a Service Manager who will act as its principal point of contact for service-related activities (reporting, support for issue resolution etc.)[13].

6.1.3 The Service Provider shall work effectively with UK Biobank, underpinned by clear defined processes, to provide comprehensive support to all users of the Platform, including UK Biobank staff and registered researchers.

6.1.4 The Service Provider shall support the ongoing service for the life of the contract, including but not limited to the provision of:

- Regular reporting on service performance;
- Defined escalation paths for issue resolution; and
- Coordination of Platform releases (including ingestion of UK Biobank Data).

6.1.5 The Service Managers shall meet at least monthly to review service performance and helpdesk activity (or more frequently as required). Such review meetings are distinct from the Executive Steering Group meetings (though they may inform) and have an operational rather than strategic focus.

[13] In the event that the UK Biobank Service Manager is not in place for the Phase 1 implementation, this role will be covered in the interim by the UK Biobank Project Manager

6.1.6 The Service Provider shall meet the Performance Levels set out in Schedule 3 of the Terms and Conditions and report against these via a regular Performance Monitoring Report.

6.1.7 The Service Provider's service reporting (which shall be made available via an online portal) shall include:

- **Performance Monitoring**: including but not limited to Platform usage (Users; compute consumption; data storage capacity, ingress and egress), service availability, Performance KPIs achievement and trends (as set out in Schedule 3 of the Terms and Conditions);
- **Capacity Planning**: a capacity plan based upon historic Platform usage and information provided by UK Biobank about anticipated growth in UK Biobank researchers, changes in the numbers of such researchers for whom UK Biobank wishes to permit Platform access, and planned UK Biobank Data releases;
- **Issue Management**: any issues observed with Platform capacity or performance in the previous period and any such risks related to the forward capacity plan, together with plans to address and/or mitigate identified risks and issues; and
- **Helpdesk Activity**: including, but not limited to, tickets raised and resolved, KPI achievement and trends; this shall cover both the Service Provider and UK Biobank elements of the Helpdesk, reported separately.

6.2 **Helpdesk**

6.2.1 The Service Provider shall provide a helpdesk capability, to support both UK Biobank staff and approved UK Biobank researchers who are Users of the Platform that will:

- provide a first point of contact for all queries and issues raised by Users of the Platform (Level 1 helpdesk), providing basic support, advising on known issues, and undertaking first level problem determination;
- escalate as necessary to Level 2 support within the Service Provider organisation all queries and issues related to use of the Platform, including capabilities, and use of the tools and pipelines it provides as standard;
- provide advice and guidance on the use of the Platform to UK Biobank staff and UK Biobank approved researchers who are Users of the Platform;
- provide advice, guidance, and technical support to Users on implementing their own tools and pipelines within the Platform; and
- escalate to Level 2 support within UK Biobank all queries and issues related to UK Biobank Access Applications and/or the use of UK Biobank Data.

6.2.2 The following Table 2 and Figure 3 show how first and second level support responsibilities are shared between the Service Provider (SP) and UK Biobank (UKB):

Domain	L1	L2
Operational issues with Platform	SP	SP
Use of Platform and native tools	SP	SP
Advice on developing/migrating tools	SP	SP
Queries and issues related to Access Applications	SP	UKB
Use and interpretation of UK Biobank Data	SP	UKB

Table 2. Support level responsibilities by domain

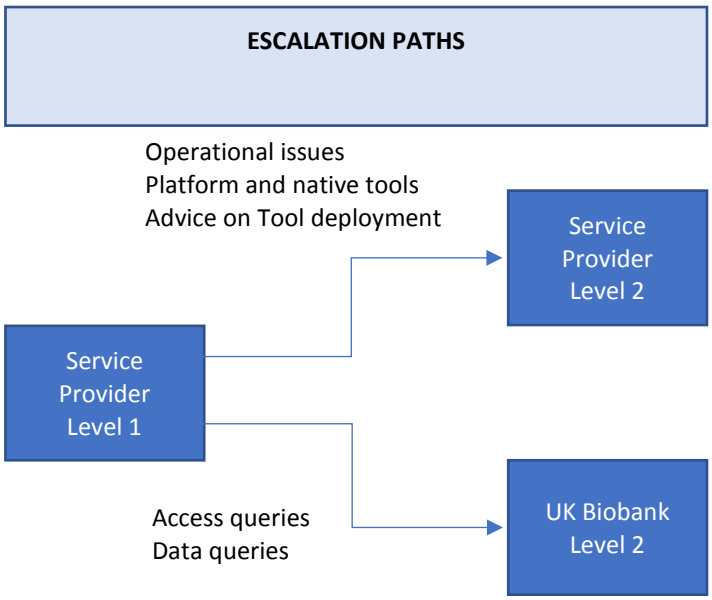

Figure 3. Escalation paths

6.2.3 Alternative means of contacting the helpdesk during operational hours shall be available, including (but not limited to) telephone, email and webchat; all queries and issues raised shall be captured within a ticketing system (which shall be accessible to UK Biobank) and tracked through to resolution.

6.2.4 Those contacting the helpdesk shall be kept informed of the progress made with resolving their query or issue, if it cannot be resolved on first contact.

6.2.5 Queries and issues raised by Users shall be reviewed to establish any underlying root cause, which shall be addressed by Platform enhancements or additional documentation (e.g. by updates to Frequently Asked Questions, the provision of additional support materials, or the creation of new tutorials).

6.2.6 The Service Provider shall support researchers in their use of the Platform, providing advice and guidance on using the capabilities the Platform provides, on appropriate tools to support their intended analyses, and on new or improved capabilities as these are made available.

6.2.7 Additional forms of self-help shall be available including the use of a community forum where users can discuss the use of the Platform and, for example, performance of different Platform tools.

6.2.8 The Service Provider shall support researchers in deploying their own tools to the Platform, providing advice and guidance to them on implementing and optimising their tools and workflows in the Platform environment - including making such tools and workflows available to other researchers using the Platform.

6.3 Documentation

6.3.1 Comprehensive documentation shall be available detailing the capabilities the Platform provides and the manner of their implementation, including best practice guidance on Platform exploitation, and Frequently Asked Questions addressing commonly encountered questions and issues.

6.3.2 The core facilities of the Platform (including the administrative tools, the analytic tools and workflows available natively, and the facilities for deploying additional tools and pipelines) shall be supported by comprehensive documentation, and this shall be supplemented by online tutorials and/or videos.

6.3.3 A data dictionary describing the UK Biobank Data available within the Platform, including a description, format, size, volume, access, availability, and semantics, shall be accessible from within the Platform.

6.3.4 All documentation and tutorials shall be accessible from the user interface, and the functionality exposed through that interface shall be supported by contextual help.

APPENDIX A – SUPPORTING INFORMATION

A.1 Managing Access Application Pseudonymised Participants Identifiers (EIDs)

A.1.1 UK Biobank maintains a number of mechanisms to ensure that all participant data released to approved researchers are provided on a de-identified basis. A key element is that each approved Access Application receives a version of data that has been keyed using a set of randomly generated pseudonymised identifiers specific to the Application (in effect, 'watermarking' the dataset).

A.1.2 It is important that this mechanism is maintained and the following examples serve to illustrate how UK Biobank considers this shall work within the Platform.

A.1.3 **Example 1 – Approved User with a single Access Application**

The numbered list in the descriptive narrative that follows corresponds to the numbers within Figure 4 below.

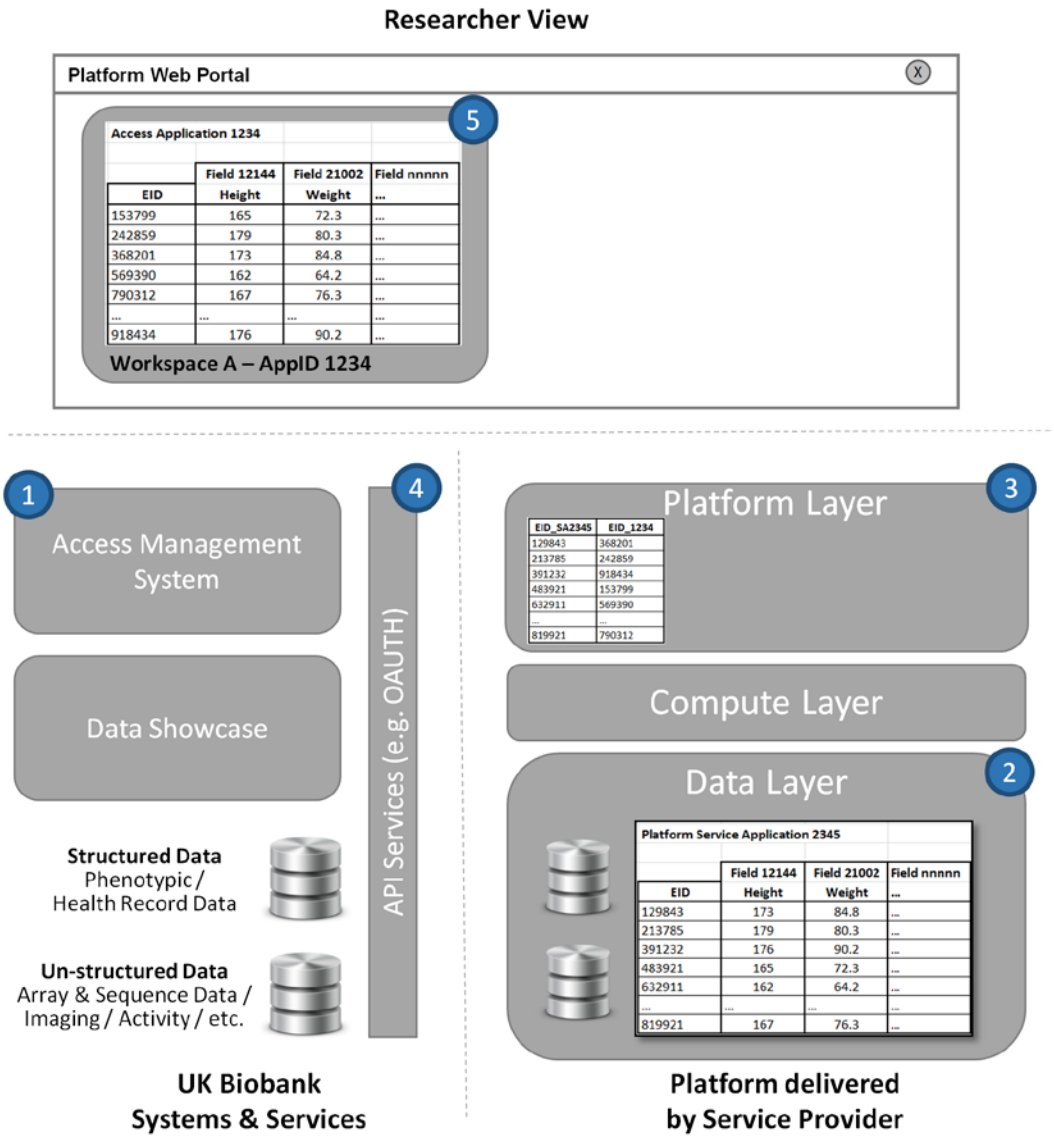

Figure 4. Managing Access Applications and pseudonymised identifiers (*Example 1*)

Example 1 - Process Flow

1. Existing UK Biobank Systems. UK Biobank hosts two key IT systems that the Platform will need to interface with as part of day-to-day operations:

- **Access Management System** (AMS) provides the ability for researchers to register to submit Access Application requests to use the resource, and manages the review process from submission to Application approval. The AMS maintains a database of researcher credentials; and
- **Data Showcase** provides a visual interface for researchers to navigate and explore the UK Biobank Data resource, and manages permissions (which data-fields researchers are able to access) and mappings (maintaining the unique set of participant pseudonymised identifiers – known as EIDs - that an Application dataset has been keyed against). The Data Showcase underpins the current delivery systems for data download.

2. Platform holding a copy of the UK Biobank Data resource. UK Biobank will regard the Platform-held master dataset as a special kind of Access Application (known as a 'Service Application'). The Service Application (depicted in the diagram as having Application ID 2345) is similarly keyed with a unique set of participant pseudonymised identifiers (EIDs). These EIDs are specific to the Platform and shall never be exposed directly to researchers.

3. Platform authentication and authorisation. For the initial implementation, UK Biobank requires that Platform access be restricted to only UK Biobank approved researchers (and where authorisation to sign-up to use the Platform has additionally been enabled within the UK Biobank AMS).

As part of the sign-up process to use the Platform, and for each subsequent Platform log-on, the Platform will need to allow Users to authenticate themselves using their UK Biobank credentials by calling services exposed by UK Biobank's API service layer.

The User may only work with UK Biobank Data within the context of a workspace. At the point of adding UK Biobank Data to the workspace, the User shall specify the Access Application ID and the Platform shall interrogate the UK Biobank API to ensure the User has the appropriate permissions. In addition, the Platform shall also obtain a copy of the mappings that link the requested Access Application EIDs to those used in the Platform.

4. UK Biobank API Services. UK Biobank will provide a number of externally callable API Web services that will allow the Platform to:

- Authenticate researchers using UK Biobank credentials;
- Check a researcher's permission to access an Application;
- Request the mapping to link the Application to the Platform Application; and
- Check which data fields the researcher's Application has access to.

Detail of the UK Biobank API is provided in Appendix A.3.

5. Researcher Web Portal View. The User may have one or more workspaces within the Platform portal. In this example, they have only one workspace within which they have requested to work with UK Biobank Data belonging to their Application with App ID 1234.

The User shall only be able to see UK Biobank Data within this workspace linked to the EIDs for their Application (and where the Platform remaps to these EIDs using the mapping file it has obtained). This approach shall extend to both structured and unstructured data (and in the latter case requires bulk data files (e.g. genome CRAMs) to be presented with remapped filenames (based on the application EIDs)).

A.1.4 Example 2 – Approved researcher with more than one Access application

Researchers may have more than one Access application with UK Biobank, and it is important that the Platform ensures isolation between these two Applications is maintained.

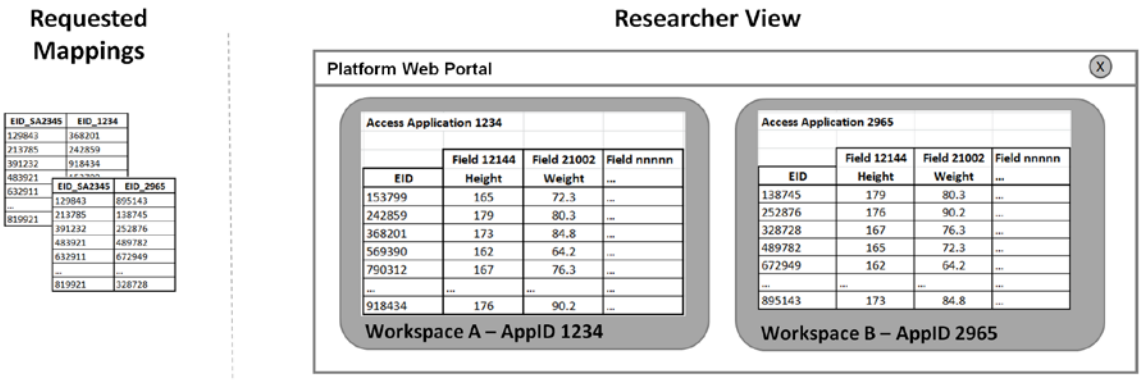

Figure 5. Managing Access Applications and pseudonymised identifiers (*Example 2*)

In this second example, the researcher has two separate Access Applications with UK Biobank. The User has created workspace 'A' to work with App ID 1234 and workspace 'B' to work with App ID 2965.

Each workspace can be linked to one (and only one) Access Application, and the User shall not be able to share UK Biobank Data between workspaces linked to different Applications. The User shall only see the appropriate pseudonymised participant identifiers (EIDs) for the Application they are working with at any one time.

In this example, the Platform will have interrogated the UK Biobank API to retrieve the mapping files for both Applications belonging to the User, to ensure that the appropriate EIDs are used for each workspace. These mapping files are static (aside from the management of participant withdrawals) and can be maintained within a central Platform repository after they have initially been called and retrieved from UK Biobank.

A.2 UK Biobank Data Specification and Access Mechanisms

A.2.1 **Phenotypic Data.** Researchers currently select the tabular data they wish to work with using Data Showcase[14] by creating a "basket" of fields for download which are keyed against the pseudonymised participant ids (EIDs) specific to their project.

The tabular data comprises up to ~27,000 distinct attributes per participant including physical measures, questionnaire responses, imaging derived measures, and other phenotypic and demographic information.

Tabular phenotypic data will be made available to the Service Provider via the existing UK Biobank Data Showcase service available to other researchers. A data basket will be created for the Service Provider Access Appliction containing all non-restricted data fields which can then be downloaded for ingestion into the Platform.

Additional information regarding the names, types and encoding meta-data related to the phenotype file is available from the Schemas section of the public UK Biobank Data Showcase at:

http://biobank.ndph.ox.ac.uk/showcase/schema.cgi

A.2.2 **Linked healthcare records.** Researchers currently access linked healthcare record data via a SQL portal, which allows researchers to interrogate the data and download the results of queries.

Linked healthcare record data will be made available to the Service Provider using the standard UK Biobank Data Portal. The UK Biobank Application associated with the Platform will be approved to access all non-restricted healthcare data.

Record level primary care data includes tables covering clinical, prescription, and registration data; they are described more fully here:

http://biobank.ndph.ox.ac.uk/showcase/showcase/docs/primary_care_data.pdf

Record level secondary care data includes hospital inpatient, psychiatric inpatient, maternity inpatient, diagnosis, operation, and delivery data; they are described more fully here:

http://biobank.ndph.ox.ac.uk/showcase/showcase/docs/HospitalEpisodeStatistics.pdf

A.2.3 **Whole Genome Sequence (WGS)**. WGS data will be made available to the Service Provider from the European Bioinformatics Institute (specifically, by downloading from the European Genome Archive, EGA).

The Service Provider will be provided with an EGA account with permissions to access the appropriate genetic data. These data may be accessed using standard EGA tools available here: https://github.com/EGA-archive/ega-download-client

WGS data includes (for each participant) CRAM, CRAI, gVCF and tbi files, together with an archive containing BQSR and QC report files.

[14] UK Biobank's online portal to present the data available for health-related research in a comprehensive and concise way, and to provide technical information for researchers considering applying to use the resource - http://biobank.ndph.ox.ac.uk/showcase/

The data package for each whole genome sample comprises (with md5 checksums):

- Raw genome data in compressed format:
  - UKBBnnnnnnn.cram
  - UKBBnnnnnnn.cram.crai
- Variant-level data in gvcf format:
  - UKBBnnnnnnn.g.vcf.gz
  - UKBBnnnnnnn.g.vcf.gz.tbi
- Ancillary files, including QC output, such as:
  - UKBBnnnnnnn_all_qc_results.zip
  - UKBBnnnnnnn.bcftools_stats.tgz
  - UKBBnnnnnnn.multiqc_report.html

where UKBBnnnnnnn refers to a sample identifier specific to the WGS assay project. When sequence data files are made available to the researcher, the filenames shall be remapped to the EIDs specific to their Application.

The EBI maintains a 20 Gbps outgoing connection to the JANET which is shared among all EBI resources. The bandwidth that would be achieved for data transfer to the Platform will be dependent upon contention with other EBI users and the Service Provider's own local resources.

A.2.4 **Whole Exome Sequence (WES).** WES data will be made available to the Service Provider from the EBI using the same tools as for WGS data. Similar to WGS data, when sequence data files are made available to the researcher, the filenames shall be remapped to the EIDs specific to that application.

A.2.5 **Array Data**. Genotyping and imputation array data in PLINK and BGEN formats will be made available to the Service Provider from either UK Biobank directly, or from the EBI using the same tools as for WGS and WES data. These data do not contain specific pseudonymised identifiers as the data format uses ancillary mapping files (FAM files) that link Application-specific EIDs to sample positioning within the data.

A.2.6 **Non-genetics data (e.g. Imaging, Activity)**. UK Biobank provides utilities for researchers to download bulk data (i. e. BLOB data such as imaging, activity, fundus images etc.). Files related to individual participants are renamed on download by these utilities to make use of the relevant project EIDs.

In the case of these bulk data, these data will be made available to the Service Provider using the standard UK Biobank ukbfetch service described here:

https://biobank.ctsu.ox.ac.uk/showcase/refer.cgi?id=644

UK Biobank's internal data delivery systems are connected to the JANET network by a 10Gbps link which is shared amongst other University of Oxford resources. The bandwidth that would be achieved for data transfer to the Platform will be dependent up on contention with other University users and the Service Provider's own local resources.

A2.7 **Other data**. There are other data types that shall be incorporated from within the UK Biobank Data resource, such as returned data from researcher Access Applications (as it is a key principle that researchers returned derived data as part of their Material Transfer Agreement). UK Biobank will wish to work with the Service Provider to consider the most appropriate way for these data to be incorporated over time.

A.3 UK Biobank Web Services API Specification

Please Note: the API specifications provided in this section represent UK Biobank's current implementation and are subject to change as UK Biobank may deem necessary.

A.3.1 Authentication

Authentication is the process whereby a platform determines the validity of a User and the domains (i.e. UK Biobank Access Applications) they are allowed to access. It is intended that UK Biobank password credentials remain solely within the UK Biobank Application Management System (AMS) at all times, so real-time communication is required between that and the Platform when sessions are to be initiated; the Platform shall never have direct access to UK Biobank User passwords.

It is assumed that the Platform will link a UK Biobank username to a chargeable entity in order that utilisation fees, such as CPU usage and additional storage, can be allocated and paid. The Platform-specific login credentials for this chargeable entity (e.g. a Platform-specific username) shall be verified at each login to guard against allowing an unauthorised person to run up large charges against someone else's account.

OAuth Authentication Service

UK Biobank has implemented (and requires the Service Provider to use) an OAuth 2.0 authentication service to enable the Platform to authenticate Users and discover key information about them. UK Biobank is providing two services:

- An authentication service utilising the OAuth 2.0 protocol; and
- An information service utilising a REST API to provide details about the User and their Access Applications.

Note that all connections to these two services are logged.

The UK Biobank authentication service implements the server-side of the OAuth 2.0 authorisation specification RFC6749 (https://tools.ietf.org/html/rfc6749) of the IETF protocol authorisation code flow. Platforms need to implement the client side of the OAuth 2.0 specification. Sections 1.3.1 and 4.1 of RFC6749 are particularly pertinent. There are 3 parties in the authorisation process:

- Server (UK Biobank);
- Client (the Platform); and
- User (an approved Researcher registered with UK Biobank).

The authentication service endpoints are (note: these are test endpoints):

- Authorisation endpoint:
  https://bbams.ndph.ox.ac.uk/aoTesting/oauth/authorize
- Token endpoint
  https://bbams.ndph.ox.ac.uk/aoTesting/oauth/token
  Scopes: applications+user

The Platform will be issued with client credentials in the form:

- clientID: XXX
- clientSecret: XXX

Prior to use, a Platform shall inform UK Biobank of its client application redirection endpoint. Please note: UK Biobank will issue client credentials to Bidders who respond to the Expression of Interest however each Bidder will need to confirm its client application redirection endpoint in order to use these services as part of the demonstration evaluation.

The Platform will obtain an access token at the end of the authorisation process, and will then be able to use the token to request further information via the information service.

The information service implements a secure REST API with two endpoints aimed at Platform clients:

- https://bbams.ndph.ox.ac.uk/aoTesting/applications/<applicationId>
- https://bbams.ndph.ox.ac.uk/aoTesting/user

The application endpoint indicates whether a given Application ID is accessible by the User. The user endpoint gives information on that User. Both endpoints emit JSON.

Simplified examples of the HTTP requests in the OAuth 2.0 protocol are provided below as a getting-started guide. The OAuth 2.0 specification RFC 6749 of the IETF is the authoritative guide and provides more comprehensive examples. UK Biobank's demonstration OAuth client: https://bbams.ctsu.ox.ac.uk/aocTesting/ is available for reference.

The primary reference for OAuth 2.0 is: https://tools.ietf.org/html/rfc6749.

Example: OAuth authentication service

Example HTTP requests in the OAuth 2.0 protocol authorisation code flow are given below.

Authorisation request

See 4.1.1 of RFC6749. The client directs the user-agent to the authentication service:

```
GET /aoTesting/oauth/authorize?response_type=code&
client_id=<clientid>&redirect_uri=<clientCallbackURI>&
scope=applications+user&state=<stat> HTTP/1.1
Host: https://bbams.ndph.ox.ac.uk
```

Authentication service authenticates the User

Login requests and responses are between the authentication service and the User and are out of scope of this guide.

Authorisation response

See 4.1.2 of RFC6749. The authentication service directs the user back to the client:

```
HTTP/1.1 302 Found
Location: <clientCallbackURI>?code=<code>&state=<state>
```

Access token request

The client exchanges the code for an access token:

```
POST /aoTesting/oauth/token HTTP/1.1
Host: https://bbams.ndph.ox.ac.uk
Authorization: Basic <credentials>
Content-Type: application/x-www-form-urlencoded
grant_type=authorisation_code&code=<code>&
redirect_uri=<clientCallbackURI>
```

where <credentials> = clientid:clientsecret base64 encoded see RFC7617.

Access token response

```
HTTP/1.1 200 OK
Content-Type: application/json;charset=UTF-8
{
"access_token":"<token>",
"token_type":"bearer",
"expires_in":3600,
"scope":"applications user"
}
```

Requests to information service for application and user data

```
GET /aoTesting/application/42992 HTTP/1.1
Host: https://bbams.ndph.ox.ac.uk
Authorization: Bearer <token>
```

```
GET /aoTesting/user HTTP/1.1
Host: https://bbams.ndph.ox.ac.uk
Authorization: Bearer <token>
```

A.3.2 Configuration

Configuration is the process whereby the participant data held by the Platform (under its Service Application) is protected and pseudonymised for use by Users (who are working as part of different Applications). Configuration information for a particular Application will only be supplied once at least one User has logged onto the Platform and successfully connected to a workspace associated with that Application.

UK Biobank has implemented a CGI-based web service to provide two types of configuration information:

- Permission information providing details of the data fields an Access Application has permission to use; and
- Mapping information providing the correspondences between the EID scheme used by the Access Application and that used by the Platform.

Configuration Web Service

Once a researcher has been authenticated via the OAuth 2.0 authentication service, the Platform shall use the configuration Web service to retrieve privilege and configuration information related to their UK Biobank username. This is located at the URL:

https://biobank.ndph.ox.ac.uk/service/info.cgi

All calls to the Web service shall include the following compulsory parameters:

Parameter	Type	Meaning	Example
prj	String	Project identifier	"ukb"
src	String	Source of caller (platform) identifier	"tender"
act	String	Action requested	"doapp"
tim	Integer	Timestamp, C time()	1142877620
tok	String	Authentication token	"tender"

In addition to the compulsory parameters, additional ones will be required according to the value of the action requested.

Valid responses begin with the characters "OK" followed by a newline.

Lines may be present beginning with a # character. These are purely comments, typically present for transient development or debug purposes, and shall be ignored when processing the body of the message.

Responses are plain ASCII text. The last 32 characters of each response are a lower-case hexadecimal representation of the MD5 checksum for the entire preceding message body (including any comment lines). This checksum shall be verified on each occasion and the response rejected if it does not match.

Please note: For the purposes of the demonstration evaluation, the service will accept requests via both GET and POST, the value of the timestamp need only be a positive integer, and all incoming IP addresses will be accepted.

Specific data documentation for details of input and outputs for the two actions that can be requested are provided below.

Permission Information

This is a service (act="fieldlist") which lists the data-field IDs that a particular Application ID is allowed to access. In addition to the compulsory requirements for the configuration web service, it also requires the parameter:

Parameter	Type	Meaning	Example
app	Integer	Application ID identifier	453

The response is a plain text string with values separated by newline characters. Contents are:

- The first row is “OK” on success;
- The second row echoes the Application ID;
- The third row contains the count N of data-fields that will follow;
- The following N rows list the data-fields to which access is granted;
- The final row is an MD5 checksum for all contents of the response prior to itself.

Below is shown the output that would be expected on querying the system for permission information related to Application ID 453 which returns only 3 allowed field IDs (5428, 18963 and 9874).

```
OK
453
3
5428
18963
9874
A123456b53A123456b53123456b53
```

If no access is permitted, then the row count is 0 and, in place of the list of data-fields, a single row will be presented containing an informative textual error message (which the vendor can use to alter the request or report to UK Biobank for assistance). The following example illustrates the output that would be expected if a query was made for non-existent Application ID 617.

```
OK
617
0
Application 617 is not known to the system
A123456b53A123456b53123456b53
```

If the first line is not “OK” then the contents are an error message explaining to the vendor the reason for failure.

The following URLs will retrieve the data relevant to example Access Applications referenced as part of the demonstration evaluation:

https://biobank.ndph.ox.ac.uk/service/info.cgi?prj=ukb&src=tender&act=fieldlist&tim=123&tok=tender&app=42992

https://biobank.ndph.ox.ac.uk/service/info.cgi?prj=ukb&src=tender&act=fieldlist&tim=123&tok=tender&app=43027

Mapping Information

This is a service (act=”eidmap”) which gives the correspondences between the EID scheme used by the Platform Application (e.g. Application ID = 11) with that used by the User’s Application enabling an Application-specific pseudonymisation mask to be applied. In addition to the compulsory requirements it also requires the parameter:

Parameter	Type	Meaning	Example
app	Integer	Application ID identifier to which map required	453

The response is a plain text string with values separated by newline characters. Contents are:

- The first row contains “OK” on success;
- The second row echoes the Platform Application ID (i.e. that associated with the Platform, which is known by the configuration web service);
- The third row echoes the target Application ID;
- The fourth row contains the count of linked ID pairs that will follow;
- Subsequent rows list the Service EID followed by the corresponding Target EID, with the two values separated by a tab character;
- The final row is an MD5 checksum for all contents of the file prior to itself.

The output below would be produced by writing to file the return value that might be generated when the platform is Application 11, the Target is Application 453 and there are only two EIDs in the mapping (such that person 5428876 in the Platform dataset corresponds to person 4309361 in the Researcher dataset).

```
OK
11
453
2
5428876 4309361
2318963 7652028
A123456b53A123456b53123456b53
```

If it is not possible to link EIDs, then the row count is 0 and, in place of the list of correspondences, a single row will be presented containing a textual error message.

```
OK
11
617
0
Application 617 is not known to the system
A123456b53A123456b53123456b53
```

If the first line is not “OK” then the contents shall be treated as an error message.

The following URLs will retrieve the data relevant to example Access Applications referenced as part of the demonstration evaluation:

https://biobank.ndph.ox.ac.uk/service/info.cgi?prj=ukb&src=tender&act=eidmap&tim=123&tok=tender&app=42992

https://biobank.ndph.ox.ac.uk/service/info.cgi?prj=ukb&src=tender&act=eidmap&tim=123&tok=tender&app=43027

APPENDIX B – FUNCTIONAL REQUIREMENTS BY PHASE

The following table lists the requirements described in this Specification and uses a MoSCoW prioritisation to indicate for which Phase they **should (S)** or **must (M)** be delivered.

Reference	Requirement	Phase 1	Phase 2
4.1	**Access and Analysis**		
4.1.1	**Web Portal and API Access**		
4.1.1.1	A User shall be able to access data and analysis services via a Web Portal	M	M
4.1.1.2	A User shall be able to access data and analysis services via Platform API(s)	M	M
4.1.1.3	The Web Portal shall be accessible via widely supported Web browsers	M	M
4.1.1.4	Only approved UK Biobank researchers shall be able to register as Users	M	M
4.1.1.6	The Platform shall integrate with UK Biobank Web Services	S	M
4.1.1.7	The Platform shall use UK Biobank Web Services to determine whether an individual is an approved researcher	S	M
4.1.1.8	The Platform shall use UK Biobank Web Services to verify that a User is part of a specific Access Application	S	M
4.1.1.9	The Platform API shall support invocation of Platform services for data access and analysis via command line tools	M	M
4.1.2	**Workspaces**		
4.1.2.1	Each Access Application shall have read-only access to a subset of UK Biobank Data	M	M
4.1.2.2	The data available to an Access Application shall be keyed using its own specific set of UK Biobank-supplied pseudonymised identifiers (EIDs)	M	M
4.1.2.3	The Platform shall use UK Biobank Web services to determine the subset of UK Biobank Data available to an Access Application	S	M
4.1.2.4	The Platform shall use UK Biobank Web services to determine specific EIDs for each Application	S	M
4.1.2.5	A User shall be able to access suitably keyed UK Biobank data for review within the context of a virtual project area or workspace	M	M
4.1.2.6	A User shall be able to access suitably keyed UK Biobank data for analysis within the context of a virtual project area or workspace	M	M
4.1.2.7	A User shall be able to access suitably keyed UK Biobank data for download (if permitted) within the context of a virtual project area or workspace	M	M
4.1.2.8	The Workspace shall control access to UK Biobank Data or other data	M	M
4.1.2.9	The Workspace shall allow these data to be analysed by Platform-provided and other tools	M	M
4.1.2.10	The Workspace shall allow these data, analysis code and/or results to be appropriately shared with Collaborators	M	M
4.1.2.11	The Platform shall regard the UK Biobank Data it holds as a read-only master dataset	M	M
4.1.2.12	Data made available through a workspace shall be via linking to the Platform's copy of the data	M	M
4.1.2.14	The Platform shall be able to manage permissions at the Access Application level	S	M
4.1.2.15	The Platform shall be able to manage permissions at the data field-level to determine which UK Biobank Data the User is allowed to access.	S	M
4.1.2.16	The Platform shall allow Users to create multiple workspaces	S	M
4.1.2.17	The Platform shall allow Users to associate an individual workspace with one (and only one) UK Biobank Access Application	M	M
4.1.2.18	The Platform shall confirm that the User is an approved researcher on the Access Application (otherwise access shall be denied) both at the time of initial association and on subsequent access	M	M
4.1.2.19	The Platform shall allow UK Biobank Data to be copied (or shared) between workspaces associated with the same Access Application	S	M
4.1.2.20	The Platform shall use the UK Biobank Web Services to confirm User permissions before UK Biobank Data are initially linked to the workspace	S	M
4.1.2.21	The Platform shall use UK Biobank Web Services to confirm User permissions on each subsequent access session through the Platform (whether via Web portal or API)	S	M
4.1.2.22	The creator of a workspace shall be able to invite other Users to join it	S	M
4.1.2.23	The Platform shall ensure that invitees are Collaborators on the same Access Application	S	M
4.1.2.24	The creator of the workspace shall be able to add charging and billing details	S	M
4.1.2.25	The creator of the workspace shall define the Users who can invite other Users to the workspace	S	M
4.1.2.26	The creator of the workspace shall define the invited Users accountable for charging and billing	S	M

4.1.3	**Interactive Analysis (including Visualisation)**		
4.1.3.1	The Platform shall enable Users to interactively explore, analyse and visualise the data	S	M
4.1.3.2	The Platform shall support a variety of user interfaces based upon standard bioinformatics research tools	S	M
4.1.3.3	The Platform shall include a Cohort Browser to support data browsing and sub-cohort definition based upon phenotypic, genotypic or other participant characteristics	S	M
4.1.3.4	The Cohort Browser shall only provide access to data the User is authorised to access from the Access Application	S	M
4.1.3.5	The Cohort Browser shall allow definition of multiple sub-cohorts based on user-defined phenotypes	S	M
4.1.3.6	The Cohort Browser shall support saving and sharing of sub-cohort definitions	S	M
4.1.3.7	The Cohort Browser it shall allow any attribute or data element to be used as a sub-cohort selection criterion	S	M
4.1.3.8	As a sub-cohort is created within the Cohort Browser the number of participants for whom data are available shall be indicated, and (where relevant and possible) an indication of the data distribution shown	S	M
4.1.3.9	The Platform shall include a Notebook capability (such as that provided by Jupyter Notebooks) for Users who may wish to develop their own code for data analysis	M	M
4.1.3.10	The Notebook capability shall include support common programming languages used for data analysis, including Python and R	M	M
4.1.3.11	The Platform shall support Notebooks that require access to distributed processing (e.g. Apache Spark) for the running of specialised tools, such as HAIL	M	M
4.1.3.12	The Platform shall support the development, upload, and deployment of other interactive applications	S	M
4.1.3.13	The Platform shall provide graphical tools appropriate for viewing genetic data and results for visual validation of genetic variants	S	M
4.1.3.14	The Platform shall provide graphical tools appropriate for viewing outputs from association studies	S	M
4.1.4	**Batch Analysis (including Tools and Pipelines)**		
4.1.4.1	The Platform shall support range of analytical use cases	M	M
4.1.4.2	The Platform shall support batch-level processing of raw genetic data to generate high quality variant call-sets	M	M
4.1.4.4	The Platform shall support phenotype analyses	M	M
4.1.4.5	The Platform shall support genotype analyses	M	M
4.1.4.6	The Platform shall support association analyses – both GWAS and PheWAS	M	M
4.1.4.7	The Platform shall support translational analyses	M	M
4.1.4.8	It shall be possible for the User to define their own pipelines which incorporate both Platform native tools and User tools	M	M
4.1.4.9	The tools and pipelines available as part of the Platform shall include those that support:	M	M
4.1.4.9.1	· standard statistical analyses such as Stata, SPSS, SAS, R (subject to appropriate licensing);	S	M
4.1.4.9.2	· common processing of genomic data (such as GATK);	M	M
4.1.4.9.3	· WGS and WES variant calling from sequence data;	M	M
4.1.4.9.4	· manipulation of variant call file outputs (such as bcftools or samtools);	M	M
4.1.4.9.5	· analysis of alignment variation;	M	M
4.1.4.9.6	· association studies, such as GWAS and PheWAS analyses;	M	M
4.1.4.9.7	· analysis of population structure and relatedness;	M	M
4.1.4.9.8	· polygenic risk scoring;	M	M
4.1.4.9.9	· image processing, for derivation of image-derived phenotypes; and	S	M
4.1.4.9.10	· machine learning.	S	M
4.1.4.10	Users shall be able to run analysis processes by invoking them directly through the user interface	M	M
4.1.4.11	Users shall be able to run analysis processes by including them in analytic pipelines	M	M
4.1.4.12	Users shall be able to run analysis processes by invoking them programmatically using the Platform API(s)	M	M
4.1.5	**Tool Repository**		
4.1.5.1	The tools and pipelines available within the Platform shall be described in, and accessible through, a searchable repository	S	M
4.1.5.2	The tool repository shall be consistent with the patterns set out in the GA4GH Tool Registry	S	M

	Service API		
4.1.5.3	Tools and pipelines that are part of the core Platform shall be optimised to make best use Platform services and underlying cloud infrastructure	S	M
4.1.5.4	Tools and pipelines shall be version controlled to support change control and reproducibility	S	M
4.1.5.5	When new versions of tools and pipelines are released as part of the Platform, access to previous versions shall be easily (and identifiably) available	S	M
4.1.5.6	Version information for the tools and pipelines used in a specific analysis shall be logged to support reproducibility	S	M
4.1.5.7	The Platform shall be extensible, with capabilities for Users to use additional tools and pipelines that are not part of the standard Platform offering	M	M
4.1.5.8	It shall be possible for Users to incorporate their own tools and pipelines using container technologies such as Docker or Singularity	S	M
4.1.5.9	The Platform shall limit any damage caused by running malicious code solely to the workspace associated with the User running it	S	M
4.1.5.10	Tools and pipelines shall be shareable with other researchers associated with the same UK Biobank Access Application, with other members of the researcher's institution, with their collaborators, or be made public	S	M
4.1.5.11	It shall also be possible for tools developed by others elsewhere (subject to appropriate licensing) to be accessed from and used within the Platform	S	M
4.1.5.12	The Service Provider shall maintain a process for the prioritisation and incorporation of additional tools based on User demand and priorities agreed with UK Biobank	S	M
4.1.6	**Automation (including Workflow and Process Execution)**		
4.1.6.1	The Platform shall support one or more standard workflow languages for pipeline definition and automation	S	M
4.1.6.2	The Platform shall include a process execution engine to run defined pipelines and workflows	M	M
4.1.6.3	The Platform shall be able to scale workflows vertically and horizontally, and where compute instances can be defined as part of workflow definitions	M	M
4.1.6.4	Process execution engine shall be consistent with the patterns set out in the GA4GH Workflow Execution Service API	S	M
4.1.6.5	The Platform must support Users of varying level of technical ability	S	M
4.1.6.6	The user interface shall follow industry good practice in usability design	M	M
4.1.6.7	The user interface shall follow established usability principles:	M	M
4.1.6.7.1	· **Visibility**: everything to complete a task shall be available and apparent when and where needed;	M	M
4.1.6.7.2	· **Feedback**: users shall be kept informed of consequences of actions, events, and progress relevant to them and their tasks;	M	M
4.1.6.7.3	· **Structure**: layout shall be organised by meaning and use and shall make sense to Users in terms of their intentions;	M	M
4.1.6.7.4	· **Reuse**: purposeful consistency shall be maintained through reuse of internal and interface components and behaviours;	M	M
4.1.6.7.5	· **Tolerance**: the user interface shall support flexibility in interactions, allowing Users to complete tasks in a number of different ways and catering for potential User errors;	M	M
4.1.6.7.6	· **Simplicity**: important and frequent tasks shall be simple and straightforward	M	M
4.1.6.8	The user interface shall enable both new and established (expert) users to accomplish tasks efficiently and with minimal interactions	S	M
4.1.6.9	Contextual help shall be provided	M	M
4.1.6.10	Contextual help shall include links to online documentation and tutorials	S	M
4.1.6.11	The user interface shall be accessible, and support standards such as the Web Content Accessibility Guidelines (WCAG)	S	M
4.1.6.12	The Service Provider shall demonstrate that it considers usability as an underlying principle within its software development lifecycle	S	M
4.2	**Storage Services**		
4.2.1	**Data Storage**		
4.2.1.1	The Platform shall provide hosting for, and controlled access to, a copy of all UK Biobank Data	M	M
4.2.1.2	The Platform shall host structured (tabular) data comprising phenotypic data and linked healthcare record data	M	M
4.2.1.3	The Platform shall host bulk (BLOB) data	M	M
4.2.1.4	The data storage system shall maintain granular access permissions to cater for scenarios where some Access Applications only have permission to use certain subsets of UK Biobank Data	M	M
4.2.1.5	The Platform shall provide tools to monitor and report on data access and usage	M	M
4.2.2	**Data Management**		

4.2.2.1	The Platform shall support UK Biobank's strict pseudonymisation approach	M	M
4.2.2.2	The Platform shall be able to ingest steady streams of data	M	M
4.2.2.3	The Platform shall be able to ingest new data types	S	M
4.2.2.4	The Platform shall be able to accept data refreshes	S	M
4.2.2.5	The Platform shall make available UK Biobank provided documentation related to data releases	S	M
4.2.2.6	The Platform shall support removal of participant-related data following notification from UK Biobank of any participant withdrawals	M	M
4.2.2.7	Data withdrawals shall encompass deletion from the Platform copy of the UK Biobank Data those individual records and files related to the participant(s) who have withdrawn	M	M
4.2.2.8	Data withdrawals shall encompass removal from shared files such as multi-sample PLINK files or pVCFs details related to participants who have withdrawn	S	M
4.2.2.9	Withdrawals shall be completed with 24 hours of notification by UK Biobank	S	M
4.2.2.10	The Platform shall allow User generated derived variables for return (together with accompanying metadata and documentation) to be copied to a UK Biobank staging area (or equivalent) within the Platform for review	S	M
4.2.3	**Data import and export services**		
4.2.3.1	Users shall be able to upload other data directly related to their Access Application	M	M
4.2.3.2	Users shall be able to undertake combined analyses including their own or other publicly available datasets where such linkage is technically feasible	M	M
4.2.3.3	Data upload and linkage shall be accomplished using industry standard tools and open standards	M	M
4.2.3.4	Industry standard tools shall be available to data download and export	S	M
4.2.3.5	All import and export of data shall be subject to appropriate levels of User authentication and access control	M	M
4.2.3.6	Users shall be able to download and export UK Biobank Data from the Platform for further analysis using their own infrastructure	S	M
4.2.3.7	The Platform shall allow restrictions to be placed upon the download or export of data	S	M
4.2.3.8	Users shall be able to download the results of their analyses	M	M
4.3	**Compute Service**		
4.3.1	The Platform shall make available a range of compute instances that can be selected by the User based on their analysis requirements	M	M
4.3.2	The Platform shall support Linux and Windows instances	M	M
4.3.3	The Platform shall support the following instance categories:		
4.3.3.1	· general purpose compute instances;	M	M
4.3.3.2	· compute-intensive instances;	M	M
4.3.3.3	· memory-intensive instances; and	M	M
4.3.3.4	· specialised compute instances (such as GPU or FPGA nodes).	S	M
4.3.4	Users shall be able to select from a standard list of available configurations in each category	S	M
4.3.5	The Platform shall suggest recommended configurations for typical analyses and use cases	S	M
4.3.6	Users shall be able to customise their own configuration for allocation of numbers of cores, memory and/or storage	M	M
4.3.7	The Platform shall be able to support analyses being run by multiple Users concurrently	M	M
4.3.8	The Platform shall be able to support transient cases of very high levels of use (e.g. when new data are released)	S	M
4.4	**Platform Support Services**		
4.4.1	**Authentication and Authorisation**		
4.4.1.1	The Platform shall include authentication and authorisation controls so that only approved researchers are able to access the data	M	M
4.4.1.2	The Platform shall build upon the capabilities of the underlying infrastructure service to ensure data integrity and availability	M	M
4.4.1.3	Users shall register with the Platform to establish a Platform account	M	M
4.4.2	**Security and Compliance**		
4.4.2.1	Platform shall process and store the data as if it were personal data	M	M
4.4.2.3	The Service Provider shall act as a Data Processor on behalf of UK Biobank	M	M
4.4.2.4	The Service provider shall deliver the Platform in a manner that remains compliant with the GDPR at all times	M	M

4.4.2.5	Service Provider shall be able to only store UK Biobank Data in particular nodes	S	M
4.4.2.6	Service Provider shall be able to promptly disable access to UK Biobank Data from a particular storage node(s)	S	M
4.4.2.7	The Service Provider shall notify UK Biobank without undue delay in the event of any security or data breach being identified	M	M
4.4.2.8	The Service Provider shall maintain regular penetration testing and software / infrastructure vulnerability assessments	M	M
4.4.2.9	The Service Provider shall be able to demonstrate that it maintains appropriate security controls (encompassing the scope of the Platform including its underlying infrastructure) through certification to ISO27001	M	M
4.4.2.10	Data shall be encrypted at rest and in transit (when transitioning outwith the Platform) using AES-256 (or equivalent) encryption	M	M
4.4.2.11	There shall be controls in place to securely remove data when it is no longer needed (for example, at the end of the service contract); and	M	M
4.4.3	**Audit**		
4.4.3.1	An audit trail shall be maintained of Platform activity, covering data access and download	M	M
4.4.4	**Charging and Billing**		
4.4.4.1	The Platform shall ensure all other costs associated with compute utilisation, storage of additional data (whether imported from external sources or derived from analyses undertaken on the Platform), and any data ingress/egress shall be borne by individual Users	M	M
4.4.4.2	The Platform shall allow limits to be set on the charges that can accrue at the workspace, User, or institution level	M	M
4.4.4.3	The Platform shall allow limits to be set by specifying a maximum set of resources (cores, memory) and/or a maximum cost per month	S	M
4.4.4.4	Resource consumption shall be logged by User/workspace	M	M
4.4.4.5	The Platform shall allow Users to be linked to organisational entities	S	M
4.4.4.6	The Platform shall allow billing across an organisation to be monitored and controlled	S	M
4.4.4.7	The Platform shall make it evident to a User what resources will be used as part of specific analyses	M	M
4.4.4.8	The Platform shall provide guidance on the costs that may be incurred in advance and/or provide the ability for Users to cap costs by stopping an analysis job if a threshold is exceeded	M	M
4.4.4.9	If there are multiple ways of servicing a request these shall be indicated and the User allowed to select the most appropriate approach	S	M
4.4.4.10	Where data are to be uploaded or downloaded by the User, an indication of the costs associated with their ingress, storage, or egress shall be provided	M	M
4.4.4.11	Users shall be required to provide billing details before consuming any chargeable Platform resources.	S	M
4.4.4.12	The creator of a workspace shall be responsible for all charges associated with it	M	M
4.4.4.13	The creator of a workspace shall be able to review resource consumption by each User accessing the workspace	M	M
4.4.4.14	Users shall be invoiced for charges they have incurred directly	M	M
4.4.4.15	The Platform shall support a variety of payment methods by which Users may settle their account, including credit card billing and institutional purchase orders	S	M

APPENDIX C - DATA REQUIREMENTS BY MILESTONE

The following table lists the data that must be available prior to each of the milestones set out in Section 5.2 of this Specification, and subsequently during the remainder of Phase 2.

Further details of the data themselves may be found in Appendix A.2 of this Specification, and volume growth over time is set out in the table in Section 4.2.1 of this Specification.

	Phase 1 Beta	**Phase 1 Production**	**Phase 2 Pilot**	**Phase 2 Production**	**During Phase 2**
Phenotypic Tabular data	All available data <1 TB	All available data <1TB	All available data <1TB	All available data <1TB	All available data <1TB
Linked Healthcare records	All available data <1TB	All available data <1TB	All available data <1TB	All available data <1TB	All available data <1TB
Whole Genome Sequence data (WGS)	Variant data for at least 10,000 participants ~75TB	All available variant (VCF) data ~1PB	All available variant (VCF) data ~1.5PB	All available (VCF, CRAM) data ~10PB	All available (VCF, CRAM) data ~12.5PB
Whole Exome Sequence data (WES)	-	-	All available variant (VCF) data ~300TB	All available variant (VCF,CRAM) data ~1.5PB	All available (VCF, CRAM) data ~1.5PB
Array data (genotyping and imputation)	All available data ~15TB	All available data ~15TB	All available data ~15TB	All available data ~15TB	All available data ~15TB
Imaging data	-	-	All available data ~280TB	All available data ~300TB	All available data ~500TB
Activity Monitoring data	-	-	-	-	All available data ~10TB
Cardiac Monitoring data	-	-	-	-	All available data ~12TB
Recruitment	-	-	-	-	All available data ~20TB