UK Biobank Limited

Procurement Name:
Managed Informatics Platform for Research Access to Data

Procurement Reference Number:
UKBB009

Procurement Procedure:
Open

Invitation to Tender (ITT)

Demonstration Evaluation

Contents

1 Introduction

This document relates to an exercise to demonstrate the capabilities of Bidder systems as part of a tender process run by UK Biobank to select a Service Provider to provide a Managed Informatics Platform service to UK Biobank.

UK Biobank would like Bidders to demonstrate the capabilities of their systems using a close analogue to the UK Biobank Data and associated services at go-live time. This document describes the specific aspects of the Platform that we would like to see demonstrated.

To assist Bidders with providing such a demonstration, a synthetic dataset and associated services have been made available and the demonstration use cases requested are framed in terms of these. Bidders may ingest the dataset described below onto their demonstration Platform and configure it for use in the same fashion they would do for live running.

It is assumed herein that Bidders are familiar with the contents and nomenclature of the live UK Biobank Data Showcase at http://biobank.ndph.ox.ac.uk/showcase/ including terms such as Application and Basket used by UK Biobank's Application Management System (AMS) and conventional download system.

UK Biobank has sought, where practicable, to keep the dataset, format and processes which form part of the demonstration similar to those which it envisages will form part of its final system when the managed service goes live.

UK Biobank reserves the right at its sole discretion to update this documentation and associated datasets during the period of the tender if unexpected problems with the supplied dataset are identified.

2 Overview

The data provided are based around the notion of the Platform being registered as UK Biobank Application 11 and serving Users "alpha", "beta" and "gamma" (user IDs 100202, 100203 and 100204 respectively) who are linked to Applications 42992 (alpha and gamma) and 43027 (beta and gamma). There are three aspects to the data:

- Static files representing a copy of the phenotype data;
- A web-service handling login authentication of users and their abilities; and
- A web-service handling Application-level configuration.

Although the files do not contain any live personal data, both they and details of the Web services, should be treated as confidential to the scope of this tender exercise. The data files are described further below in Section 3 of this document.

The Web services are described further in Appendix A.3 in the Specification document. In addition, UK Biobank will provide additional API documentation with more details (available in the same location as the demonstration synthetic dataset described below).

The platform system must provide separate workspace areas related to Applications 42992 and 43027. Within workspaces, all UK Biobank supplied data must be aliased according to the identifier scheme (EIDs) associated with the mappings obtained from the UK Biobank information service. In displays and outputs, categorical values should be interpreted using definitions available from the online UK Biobank Showcase coding lists in the schema.

During the demonstration additional usernames and Applications will be introduced to explore the registration process for new users. Note that we would not expect the system to configure itself in real-time to support new Application IDs.

3 Synthetic Data for Demonstration Evaluation

The dataset for the demonstration exercise is an analogue of the real UK Biobank dataset, containing synthetic values but preserving most of the richness of the actual data.

The demonstration dataset consists of the following elements

- Files containing simulated phenotype information equivalent to a user selecting all data-fields in the UK Biobank showcase;
- Files containing simulated genotype SNP information;
- Files containing simulated medical records information; and
- A set of roughly 10,000,000 individual files analogous to the UK Biobank 'bulk' blob-store repository which holds genetic, imaging and other per-person unstructured data.

All phenotype files described here are available for download from the following website:
https://biobank.ndph.ox.ac.uk/platform_tender/

Please note that the datasets for the demonstration are not internally consistent – for instance they may contain reports of prostate cancer linked to female participants, medical events after death or dates of disease without corresponding diagnoses. This is a consequence of the pseudo-random nature of their generation and will not affect the demonstration exercises required to prove the target outcomes.

3.1 Tabular files

This corresponds to the information issued to researchers via the standard UK Biobank mechanism of creating a Showcase basket then downloading the extracted results. It is a plain ASCII text file containing variables separated by tabs across columns and new-lines between rows. The first column of each row will be the participant identifier (EID) associated with Application 11. The dataset contains approximately 27,000 columns by 600,000 rows. A header row will be included containing "EID" followed by the data column names using the convention of

FieldID – InstanceId ▪ ArrayID

Because of the size of the dataset, the data is being delivered as a set of 23 files which each contain a non-overlapping subset of the whole, split into groups of fields (i.e. all persons are present in every field).

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

3.2 Genotype Files

This contains the SNP information produced by the UKB genotyping chip, with meta-data available from http://biobank.ndph.ox.ac.uk/showcase/schema.cgi?id=15. The dataset contains approximately 600,000 "rows" of 840,000 columns each. The data are supplied as a dictionary file plus a series of data files.

The dictionary file "gene_dic.dat" is a 7-column tab separated file, containing the columns:

- Affymetrix ID (integer)
- Chromosome ID (1-2 chars)
- Index (integer, zero-based) along data row
- SNP variant 0, e.g. "T T"
- SNP variant 1, e.g. "T T"
- SNP variant 2, e.g. "T T"
- SNP variant 3, e.g. "AGC G"

If there are less than 4 variants for a particular SNP then unused entries are blank (i.e. empty quotes).

The data are supplied as a series 26 of files named "rand_chr*.dat.gz", each containing the SNPs for a single chromosome across the whole cohort and accompanied by rand_chr.md5 giving the MD5 checksums for the uncompressed files. Data are provided as plain ASCII text, without spaces/tabs, with each line having the format:

eeeeeee ABCD….Z

where "eeeeeee" is the 7-digit EID and A….Z are the index values for the SNP variants in the same order as specified in the dictionary file.

To illustrate, the first 3 lines of rand_chr21.dat begin

1707540 312202120
7843592 302112201
5903945 312201202

In the first row, person 1707540 has value=0 for the 5th SNP (index=4). Referring to gene_dic.dat one finds the line:

52233461 21 4 "C C" "0 0" "T C" ""

which means that for the SNP with AffyID 52233461, person 170540 has genotype "C C".

* The data format supplied here corresponds to the internal format used by UK Biobank to service basket requests (typically someone asking for a few dozen SNPs). UK Biobank additionally holds the raw genotype and imputation data in PLINK / BGEN format files and the platform would also be expected to ingest these directly and make them available to researchers in their native formats – though not a part of this demonstration exercise.

3.3 Medical Records File

This is plain ASCII text containing variables separated by tabs across columns and new-lines between rows. It corresponds to the gp_clinical table accessible to registers via the portal on the Showcase website. The data is split into 6 files which together contain approximately 400,000,000 rows of 8 columns each. The columns in the files are:

- EID, 7-digit integer
- data provider, integer
- event date, date in ANSI yyyymmdd format
- read2, char(8) coding
- read3, char(20) coding
- value1, char(1024) text
- value2, char(1024) text
- value3, char(1024) text

Encoding maps for the read2 and read3 columns may be found at:
https://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=592.

3.4 Bulk Files

This is a collection of ~N,000,000 files which simulate the UK Biobank bulk file repository. The file contents however are unrelated to those found in the live system and also much smaller to facilitate download and handling in a reasonable interval (their presence in the demonstration is primarily to test the ability of the platform to handle and pseudonymise such a collection rather than to show type-specific analysis pipelines or the ability to handle large volume of storage).

Files are named according to the standard UK Biobank download convention of:

FieldID_InstanceID_ArrayID_EID.type

* When the system goes live, the platform will be expected to fetch files from the UK Biobank archive using the standard ukbfetch service under its own Application ID. For technical details related to this see:
https://biobank.ctsu.ox.ac.uk/showcase/refer.cgi?id=644.

4 Approach to the Demonstration Evaluation

As part of the tender evaluation and after Bidders have made their submissions, UK Biobank will arrange individual demonstration evaluation sessions with each Bidder. The demonstration sessions will take place during the period set out in the ITT Volume 1 and will be undertaken as a video conference with support for screen sharing. All demonstration evaluation sessions will be recorded.

Each session will last for up to 3 hours and will take the format of the Bidder demonstrating how their Platform addresses the use cases set out in the following Section 5, with time at the end for any further questions from UK Biobank. Anticipated timings for each section, and reference to the relevant section of the Specification for each use case, is provided.

UK Biobank anticipates that no more than 4 representatives from each bidder will need to take part in the demonstration, and no preliminary presentation is required.

UK Biobank expects the Bidder to utilise the synthetic dataset provided, together with linkage to the technical web services being made available, in order to fully demonstrate the use cases. Each use case demonstration will be evaluated and scored between 0 and 10 using the scoring framework and weighting set out in the ITT Volume 1.

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.

5 Demonstration Evaluation Use Cases

5.1 Security and Control (~40 minutes)

5.1.1 User on-boarding (Specification 4.1.1)

Demonstrate how a new User (who is already known to UK Biobank) registers with the platform. New usernames will be provided for this during the demonstration.

5.1.2 Logon (Specification 4.1.1)

Demonstrate that the test UK Biobank User accounts are able to login, via OAuth, using their UK Biobank credentials.

5.1.3 Access control (Specification 4.1.1)

Demonstrate that variants on the UK Biobank username+password combinations other than those supplied fail to gain entry to the platform.

5.1.4 Workspace Creation (Specification 4.1.1)

Demonstrate how a User creates a new workspace and links it to a UK Biobank Access application.

5.1.5 Workspace segregation (Specification 4.1.2)

Show that Users can only link workspaces to an Access Application with which they are associated as defined by the UK Biobank OAuth authentication service.

Note that the passwords and access rights may be changed during the course of the final interactive demonstration evaluation to detect hard-wiring of particular behaviours in the platform UI.

5.1.6 Data segregation (Specification 4.1.2)

Demonstrate that Application workspaces are able to access all the data (and only the data) defined by the permissions returned by the UK Biobank information service (action “fieldlist”).

5.1.7 Workspace management (Specification 4.1.2)

Show how groups of users on the same Application may have both private and shared spaces. Demonstrate how to share a workspace and the enforcement that all owners must be linked to the UK Biobank Application associated with it.

5.1.8 Activity Logging (Specification 4.4.3)

At the end of the demonstration, display any logs or audit trails of the activity recorded during the user sessions. If possible (working as the platform owner), export these records for external analysis.

5.2 Administrative Processes (~20 minutes)

5.2.1 User Assistance Mechanisms (Specification 6.2, 6.3)

Show the communication mechanisms whereby users of the platform may request and be given support.

5.2.2 Resource (Specification 4.3)

Show how Users assign resources (e.g. compute cores and RAM) to perform tasks.

5.2.3 Billing (Specification 4.4.4)

Show how billing for usage is carried out and the mechanism for linking it to platform accounts.

5.2.4 Financial Management (Specification 4.4.4)

Show what, if any, facilities exist to help users manage their financial outgoings - e.g. automatic caps and brakes, reporting mechanisms.

5.3 Data Management (~40 minutes)

5.3.1 Dataset selection (Specification 4.1.3)

Demonstrate a graphical interactive method (e.g. a “phenotype browser”) allowing users to identify and construct sets of participants who share particular characteristics (for instance being female, born before 1950, having 3 GP records and possessing a particular SNP). Save this set of participants for use in subsequent analyses.

5.3.2 Data completeness and visibility (Specification 4.1.2)

View specified items of data, including the contents of bulk files, for an individual person using their relevant Application-specific identifier (EID) and UK Biobank field ID and instance/array indices (action “eidmap”).

5.3.3 Foreign data incorporation (Specification 4.2.3)

Upload (working as a logged-in User) an additional file of tabular data and incorporate it alongside the existing data in a research-specific workspace. An example data file will be provided by UK Biobank at the start of the demonstration and will use EIDs matching the Application into which it is to be uploaded (as may be appropriate).

5.3.4 Non-GUI interfaces (Specification 4.1.1)

Demonstrate non-GUI (i.e. “command line”) interface(s) for accessing data. We would be pleased to see multiple methods available.

5.4 Analysis (~30 minutes)

5.4.1 Simple statistics (Specification 4.1.4)

Run a simple statistical analysis in which multiple columns of phenotype, medical record and SNP data (to be specified by UK Biobank during the demonstration) are combined. For instance, pairwise correlations and grouped counts using phenotype characteristics.

5.4.2 Tools Repository (Specification 4.1.5)

Show how a User may search the Tools Repository. Show how a user may add their own software to the Tools Repository.

5.4.3 Pipeline construction (Specification 4.1.4)

Construct a simple pipeline combining the contents of a bulk file with the value of a non-bulk variable to output both non-bulk variables and additional bulk files.

The bulk files supplied by UK Biobank for this demonstration each contain a single integer in plain ASCII text. The pipeline test will require opening the files, reading their values, then doing some simple numerical operation using these.