www.databrew.cc

BOHEMIA data management plan

Executive summary:

The Bohemia project has complex data needs. For the sake of simplicity, reproducibility, security, costs, and real-time analyzability, a centralized and consolidated approach to the management of the entire data lifecycle is required. This document describes a basic plan for the management of the data lifecycle, with a focus on the value added through a consolidated and reproducible approach.

Data collection:

Data needs to be (i) collected, (ii) validated, (iii) stored, (iv) accessed, (v) visualized, (vi) analyzed, and (vii) disseminated. DataBrew intends to manage the entire data lifecycle using custom-built, open-source software tools so that the project managers and researchers can focus on the public health aspects of the project, and not on its technological aspects.

The data collection application:

Data will be collected through a custom-made "Bohemia" Android application, running the open source OpenDataKit format, designed to run on both tablets and mobile phones. The application, will also have a browser-based component for use through a standard desktop.

Data entry:

There are three modes of data entry, listed here in order of frequency:

1. Tablet-based fieldworker / data entry clerk uses the "Bohemia" application to follow a questionnaire with explicit survey flow and validation protocol (see "Validation" section). (95% of cases)
2. Site manager uses the browser-based application to enter edge case data points or batch-processed entries. (4% of cases)
3. Centralized database manager manually enters or modifies entered data at the request of a site manager. (<1% of cases)

If required, translation/ingestion scripts will be written to port data from currently existing databases to the centralized database. However, this approach is not recommended, since it would mean not benefitting from the validation and control checks implemented by DataBrew (see "Validation" section).

Language issues:

The "Bohemia" application's front-end will be in the language of the user. Back-end translation tables will be devised for every question and answer category. Upon upload to the centralized server, a merge-and-replace with the respective language table will translate all questions and answers to English. Original-language responses will be stored separately. The centralized database will be in English, but can be translated to any local language (even when data

collection was in English) through a replace-and-merge operation. All language issues will be managed by the centralized database manager (CDM).

On-device storage and upload:

Data entry will not rely on internet connectivity. Rather, entries are stored locally on the data-entry device as "packets". When the device achieves a wifi connection with a pre-authenticated network, the packets are sent to the centralized server. Upon full data upload, a confirmation is sent from the server to the device, at which time the already uploaded data is removed from the device. For more details on storage, see "Storage and security".

Validation:

Control flow:

The "Bohemia" application will have explicitly engineered survey control flow. Questions contingent on previous responses will be shown only when applicable.

Individual validation checks:

The "Bohemia" application will employ three levels of validation: (i) confirmatory, (ii) prohibitory, and (iii) iterative. Confirmatory validation consists of requiring the person entering data to explicitly confirm an entry if deemed to be unlikely. For example, when a person under the age of 18 has a "number of children" variable of greater than 0 (unlikely, but certainly not impossible), an additional "confirmatory" check is performed to ensure that the entry was not erroneous. Prohibitory validation consists of explicitly forbidding certain entries, or requiring the approval (via a unique password) of a manager. For example, a person previously entered in the database as "male" cannot have greater than 0 pregnancies. Iterative validation consists of allowing for prohibited responses *only* if the prohibition condition is removed. For example, to enter a new malaria case for on June 1st for a person previously entered as having died on May 1st is allowed, but requires a modification of previously entered data (ie, the person either did not die or died later). Another example: if the tablet device is detected via GPS to be at a certain location, but data entry is being performed for someone who has previously been registered as residing elsewhere, a modification of the person's residence is required for further entry. This "iterative" data entry process is tracked at each iteration, so previous states can be restored by the CDM in case of error.

Aggregate validation checks:

Certain data entry events are expected at the individual level, but deemed improbable when in high frequency. For example, it may not be uncommon to have a female diagnosed with malaria. However, the consecutive diagnosis of 10 females and no males suggests potential data entry error ($p < 0.001$). These checks will be performed daily at the aggregate (ie, site) level, rather than in the data collection device. When statistically improbable occurrences arise, the site manager will automatically be prompted to confirm or correct.

Storage and security:

Storage:

A centralized PostgreSQL database will be employed via a AWS RDS (Amazon Web Services Relational Database Service). Backups (via SQL "dumps") will be generated nightly via a cron script running on an EC2 (Elastic Compute Cloud) instanced, and stored on a remote server via S3 (Simple Storage Service) as well as a local back-up server . Data tables will consist of one table per survey form, in addition to translation tables.

All sites will have access to their own data, which will also be stored on site in whichever CRM is employed by each site.

Security:

Local site managers will be given read-only SSH access to the database, and will first undergo a training regarding data security and confidentiality. If additional levels of security are required, appropriate security groups will be created by the CDM. The CDM will connect to the database using private-public keypair authentication. PHI (protected health information) such as birthdays, names, and exact coordinates and addresses will be encrypted/hashed. Hash tables for decryption will be stored separately and only be accessible to the CDM.

Access:

The three AWS services employe dy DataBrew (RDS, EC2, and S3) are industry standards and are used by some of the largest technology companies. They each come with "out-of-the-box" state of the art security measures, such as encryption via SSL, IP whitelisting, etc., while also allowing for custom security protocols. Access to any of the three services (the databases themselves, the virtual machine running the back-up scripts, or the back-up data dumps) will require multi-factor security authentication. Only authorized study managers will be given access, and data access and use will be monitored to ensure compliance with study protocol, ethical norms, and legal restrictions.

For authorized users, there will be two main points of access: (i) via CLI (command line interface) tools for tech-savvy users (such as the S3's RESTful API or the psql command utility for RDS) and (ii) a browser layer which allows for basic querying via a custom web application built by DataBrew. To avoid SQL injection or unauthorized acccess, this web application will be relatively straightforward, allowing the user to pick from a myriad of drop-down options so as to ensure successful and coherent data entry on-site, as well as to monitor data entry clerks.

Visualization (real-time dashboards):

Dynamic reporting:

By consolidating all data into a centralized location, dashboards and dynamic reports can be written *a priori* and updated in real time, connecting to the database via SSL. Two types of dynamic reporting are envisioned: (i) a process-oriented dashboard and (ii) a content-oriented dashboard.

A process-oriented dashboard is an up-to-date collection of charts, tables, and tests which detail the current status of data collection. For example, a process-oriented dashboard may show the number of certain data uploads by site location, number of manual corrections, time since last upload. It is most useful in ensuring that all data collection processes are going according to plan.

A content-oriented dashboard is an up-to-date reflection of the *content* of the data collected. For example, it might show a chart with real-time cumulative malaria incidence broken down by geography, etc.. It is most useful in monitoring results in real time so as to begin the processes of analysis and inference, *before* the final dataset is assembled.

Analysis:

During the study:

DataBrew will assist researchers to build "scripts" in Rmarkdown, iPython Notebook, Shiny, or Flask which automatically run certain content-oriented analyses which go beyond simple counts. For example, a researcher may need to know the most up-to-date estimated odds ratio for infection as a function of certain risk factors. This can be built into a dynamic report which both (a) runs with the most recent data and (b) allows for interactivity (such as the inclusion or exclusion of certain risk factors via a drop-down menu, etc.). The purpose of these analyses will be to inform the study managers, identify problems quickly, and allow for the planning of contingencies in the case of unexpected or adverse results.

After the study:

Given that DataBrew will lead the data lifecycle management during the study, it can be expected that its role after the study - in terms of data management and analysis - will be important. DataBrew will help researchers access and process data for specific research deliverables, as well as provide statistical consulting so as to ensure that data extracted and analyzed for the purpose of academic publishing is coherent and reflective of the realities of previous steps in the data lifecycle.

Dissemination:

It is expected that the main medium of research findings' dissemination will be academic journals. That said, DataBrew will work with Bohemia researchers to ensure that knowledge products which go beyond the narrow scope of academic journals is disseminated to the public. These knowledge products include (but are not limited to): white papers on a certain issue (web hosting and access managed by DataBrew); interactive visualizations and maps; pre-print web hosting; online "toolkits" and information packets regarding research findings; etc.

DataBrew will also work to ensure that the funders, the international global health community, WHO/Unitaid, and other stakeholders in the study's content and findings have access to the

most up-to-date study findings, as well as the supplementary materials which contexualize those findings. DataBrew will provide guidance on researchers regarding both (a) how to comply with privacy concerns as well as (b) how to comply with open-access and reproducibility concerns. Finally, DataBrew will manage and comply with all requests (from authorized parties) for data access (in its raw or processed form).