**The Ark: Work Plan**

**Core improvements and new functionality**

*Streamlined phone and address UI - ARK-1153*

[*https://the-ark.atlassian.net/browse/ARK-1153*](https://the-ark.atlassian.net/browse/ARK-1153)

Currently, the Phone and Address tabs are separate, and each follows the standard UI pattern of being searchable. This seems like overkill – it’s unlikely that a single person will have so many phone numbers or addresses that a search function is needed. Furthermore, the number of subject-level tabs is high, and this is an opportunity to reduce that number.

A better approach is to merge the Phone and Address tabs into a single “Contact” tab, and eliminate the search functions to streamline the UI.

*Subject contacts ARK-1154*

[*https://the-ark.atlassian.net/browse/ARK-1154*](https://the-ark.atlassian.net/browse/ARK-1154)

Some studies have expressed an interest in maintaining information about contacts of subjects, i.e. people who may or may not be directly involved in the study and have some affiliation with a subject - friends, housemates, work colleagues etc.

Subject contact information can be added to the merged Phone and Address screens (see above). This extension will require a new “Subject type” standard demographic field with values “Participant” (default) and “Contact”. The default subject listing should show “Participant” subjects by default.

*File-system based file attachments*

[*https://the-ark.atlassian.net/browse/ARK-1155*](https://the-ark.atlassian.net/browse/ARK-1155)

The Ark can store file attachments for study components and correspondence, and as “standalone” attachments (the subject-level “Attachments” tab). These are stored in the relational database schema as BLOBs, meaning that the main database tends to grow rapidly as file attachments are inserted. This in turn has implications for the disk storage holding the database, and a negative impact on efficient database backups.

A better approach is to rework the file attachment code so that the main database stores only meta-information about the files, with the files stored on a standard file-system location that is specified via a configuration parameter. In this way, the mount point for file attachments is decoupled from the database storage, and backups of the Ark’s database will not contain large amounts of BLOB data.

A standalone application will be required to migrate existing study attachments from the in-database BLOB scheme, to the new on-file-system implementation.

*Demographic custom fields – field grouping*

[*https://the-ark.atlassian.net/browse/ARK-1156*](https://the-ark.atlassian.net/browse/ARK-1156)

An unlimited number of demographic custom fields can be setup for a study; these fields appear in the “Study-specific demographic data” tab. However, the number of fields can be large, and presently the fields are simply displayed in one long, paginated list.

A better approach is to allow users to configure question *groups* (and perhaps subgroups) so that questions in a group are displayed on the same screen. The user will have UI elements to quickly navigate between question groups.

*Phenotype module – question grouping*

[*https://the-ark.atlassian.net/browse/ARK-1157*](https://the-ark.atlassian.net/browse/ARK-1157)

The field grouping solution developed for demographic custom fields is also applicable to phenotype datasets. Field grouping is even *more* important in the phenotype module since the number of fields in a questionnaire is often very large and naturally divided into sections and subsections. For effective day-to-day use of the phenotype module, quick navigation between question groups is essential.

*Change tracking for demographic data*

[*https://the-ark.atlassian.net/browse/ARK-1158*](https://the-ark.atlassian.net/browse/ARK-1158)

The Ark currently tracks changes to study-level and study component-level *consent*, but no other changes to subject-level demographic data. An improvement is to implement change tracking for the majority of standard demographic fields, in addition to correspondence and all custom fields. The user will be able to invoke a visual change log on a per field basis, and export this log to a CSV formatted file. Logging should be recorded for changes made via the web UI *and* data imports/updates that occur via the bulk loaders. We will also need to consider how changes may exportable via the Data Extraction module.

Change tracking is a very common suggestion that arises during Ark demos.

*Change tracking for phenotypic data*

[*https://the-ark.atlassian.net/browse/ARK-1159*](https://the-ark.atlassian.net/browse/ARK-1159)

Several studies have expressed a requirement to update phenotype datasets and review changes on a day-to-day basis. The change tracking, display and export functionality developed for demographic fields is also applicable to phenotype data fields. A consistent solution should be applied to both demographic and phenotype modules.

*Additional export formats*

[*https://the-ark.atlassian.net/browse/ARK-1160*](https://the-ark.atlassian.net/browse/ARK-1160)

The Data Extraction module currently produces exports in a single, fixed format: CSV. Users of analysis packages such as Stata will benefit from the addition of new export file formats.

REDCap is capable of exporting data in a variety of popular file formats; we should model our export extensions on REDCap, and aim to match its export capabilities.

*Inter-field logic for data extraction filters*

[*https://the-ark.atlassian.net/browse/ARK-1161*](https://the-ark.atlassian.net/browse/ARK-1161)

Currently, the inter-field logic between data extraction filters is a logical AND operation. This will be generalised to allow the user to specify different logic operators (e.g. OR, NOT) in addition to precedence by means of (potentially nested) bracketing.

*Data extraction – de-identification*

[*https://the-ark.atlassian.net/browse/ARK-1162*](https://the-ark.atlassian.net/browse/ARK-1162)

Study managers often have the requirement to provide *de-identified* data to researchers. New extensions to The Ark will allow de-identification in the Data Extraction module.

The de-identification process will substitute subject UIDs with configurable de-identified UIDs, and internally record the linkage of Ark subject UIDs with the de-identified equivalents. The system will be capable of exporting as CSV the mapping of Ark subject UIDs to de-identified equivalents. The de-identifier code may also restrict the export of inherently identifying data such as contact details.

*Web UI - de-identification*

[*https://the-ark.atlassian.net/browse/ARK-1163*](https://the-ark.atlassian.net/browse/ARK-1163)

In addition to de-identified exports, the security model should be extended to provide an access level that hides identifying information at the web UI level. If assigned to a user, this level of access will automatically hide identifying data such as the first and last name fields, the Phone and Address tabs, and perhaps a selection of other fields that are standard to the Demographic tab.

*Skip logic for demographic custom fields*

[*https://the-ark.atlassian.net/browse/ARK-1164*](https://the-ark.atlassian.net/browse/ARK-1164)

This extension will allow a user to specify conditional logic (i.e. “skip” logic) for Demographic custom fields. Skip logic determines whether a field is shown or hidden based upon the values of other fields, which may also be custom fields, or standard demographic fields. Conditional logic conditions will be configurable via the web UI.

The implementation of skip logic may be based on the elegant solution developed by Limesurvey. Limesurvey allows skip logic conditions to be built up with an implicit AND operation between each condition. A condition can be set using a variety of equality operators (equal to, not equal to, less than, etc.) matched against predefined values (encodings) for another field, the stored value of another field, constants, and regular expressions.

*Skip logic for phenotype datasets*

[*https://the-ark.atlassian.net/browse/ARK-1165*](https://the-ark.atlassian.net/browse/ARK-1165)

The skip logic solution developed for demographic custom fields is also applicable to phenotype dataset definitions. Since there are typically more fields in a phenotype dataset than there are demographic custom fields, skip logic for the phenotype is even *more* important in this setting. Coupled with question grouping (see above), reducing the number of fields visible to only those which are relevant (based on the current values) will markedly increase the useability of the phenotype module.

*Integration of work tracking and correspondence*

[*https://the-ark.atlassian.net/browse/ARK-1166*](https://the-ark.atlassian.net/browse/ARK-1166)

Currently, the work tracking module is independent of the subject-level correspondence functionality. This extension involves integrating the working tracking module with correspondence tracking so that bulk correspondences (mailouts, telephone campaigns, etc.) can be easily created and mapped to work requests/billable items.

*LIMS bulk import and process/aliquot/allocate*

[*https://the-ark.atlassian.net/browse/ARK-1167*](https://the-ark.atlassian.net/browse/ARK-1167)

Discussions with **life**pool have revealed that biospecimen import and processing in a lab often operates over relatively large numbers (10s, 100s) of samples at a time. The current LIMS functionality only facilitates biospecimen creation on a one-by-one basis via the UI, requiring several clicks each time. Similarly, processing and allocation of these samples is currently only possible on a sample-by-sample basis. This means that the LIMS module is time-consuming and laborious to use when bulk importing, processing and allocating samples.

A solution may be to introduce a dynamic wizard in which the user can specify a pattern for creating a batch of *n* initial biospecimens, followed by a series of processing/aliquoting steps. Ideally, the wizard should also permit inventory allocation of biospecimens each step, however this would probably require heuristics for The Ark to perform the allocation.

*LIMS “shopping cart”*

[*https://the-ark.atlassian.net/browse/ARK-1168*](https://the-ark.atlassian.net/browse/ARK-1168)

Researchers typically request batches of biospecimen samples according to a set of selection criteria (e.g. “0.1ml blood DNA from participants under 40 with colorectal cancer”). Whilst The Ark’s LIMS currently supports biospecimen transactions on a sample-by-sample basis, there is no mechanism to select a batch of biospecimens, optionally process them prior to “checkout”, automatically record the transaction for each, and track checkout requests to see which biospecimens were involved, how much of the biospecimens was taken, which researcher received the samples, etc.

In commercial LIMS packages, batch biospecimen requests are often handled using a “shopping cart” approach. The Ark may follow a similar approach, or opt for a wizard style of workflow instead.

*Integrated cancer risk modelling*

--skipping

This improvement involves using the Harvard Risk Service and the forthcoming BOADICEA web service to integrate cancer risk modelling into The Ark. This could be implemented as a new subject and/or study level “Risk” tab that draws upon demographic, phenotypic and pedigree data to form messages that are sent to risk services. The UI should provide a means for the user to configure the risk assessment (e.g. the model used, proband age, etc.), obtain risks for a batch of subjects, and export results. Ideally, the web UI will also provide risk visualisation functionality, with the ability to export visualisations as images or PDF documents.

*Web services layer*

*– skipping- could be very relevent – but Thilina may want this as part of PhD.*

Interaction with The Ark is currently only possible via the web UI or direct access to the database backend (generally, not recommended). Consequently, integrating The Ark with other data management and analysis tools is indirect and non-trivial.

A web services interface will solve this problem by providing a standard interface for external applications to interact with The Ark. The interface should facilitate data loading to The Ark, querying, and data export. Some work will be required to develop a standard interface that both development teams agree upon.

*SPHINX Ark upgrade*

[*https://the-ark.atlassian.net/browse/ARK-1169*](https://the-ark.atlassian.net/browse/ARK-1169)

The SPHINX instance of The Ark is currently the v1.1.1c release, which lacks the pedigree module, in addition to a range of bug fixes and small usability improvements. To upgrade the SPHINX Ark instance, a thoroughly tested, tagged release is needed. Scripts to update the database backend from v1.1.1c to the new release are also required.

*Seoul National University (SNU) instances*

[*https://the-ark.atlassian.net/browse/ARK-1170*](https://the-ark.atlassian.net/browse/ARK-1170)

Two Ark instances will be installed at SNU: one to host the Korean twin registry data on a standalone server (Ubuntu 12.02LTS), and another to host other datasets on a multi-purpose server (CentOS). Both instances will only be accessible by SNU researchers and the Ark development team.

As per SPHINX, we require a tested and tagged release to deploy on the SNU servers. Analysis and planning for the migration of Korean twin registry data is also required.

*User Documentation*

[*https://the-ark.atlassian.net/browse/ARK-1171*](https://the-ark.atlassian.net/browse/ARK-1171)

Very little user documentation exists for The Ark. This is detrimental to the adoption of the system, and typically means that in-person training is necessary, even for basic usage. A comprehensive user guide will also help provide a set of “smoke tests” for new release candidates.

*Internationalisation/localisation*

[*https://the-ark.atlassian.net/browse/ARK-1172*](https://the-ark.atlassian.net/browse/ARK-1172)

At the moment, The Ark does not support localization/internationalization. There are two aspects to this job: (i) translations of the user interface and its messages, and (ii) the ability to store data values in languages other than English. The lack of localisation and storage of data in foreign encodings is a stumbling block for users such as Seoul National University, South Korea.

*Modernised UI style*

[*https://the-ark.atlassian.net/browse/ARK-1173*](https://the-ark.atlassian.net/browse/ARK-1173)

A few researchers have commented that The Ark’s graphic design could look more modern and stylish. This task will require a graphic designer/web programmer to create a new look and feel. Since no one on The Ark development team is a graphic designer, this task may be best outsourced to a contractor with a demonstrated track record.

*Wicket upgrade*

The Ark uses Apache Wicket 1.5.8, which is rapidly becoming outdated; the latest stable release is 6.14.0. To move in step with Wicket development, we should migrate The Ark to Wicket 6.x. This job will require thorough testing of the migrated version prior to tagging a new release.

*Publications*

* One-page overview of The Ark
* Full-length journal article on The Ark, modelled on the REDCap paper
* Thilina’s PhD publication plan:
  + Web-based pedigree modelling and visualisation
  + A “Big data” approach to GWAS data management
  + A parallel programming approach to GWAS analyses
  + Data management and analysis using SPARK: a case study using the Twins and Sisters study
  + A systematic review of GWAS software systems
  + SPARK: a supercomputer pipeline for The Ark

*Promotion, acquiring new studies, and cost recovery*

[*https://the-ark.atlassian.net/browse/ARK-1175*](https://the-ark.atlassian.net/browse/ARK-1175)

In the Victorian instance, The Ark is lacking a breadth of studies and users. This is partly due to certain key functionalities lacking from the system (see above), and partly due to a lack of active promotion and a well-defined cost-recovery scheme.

Once The Ark has reached a greater state of maturity, we need to actively promote the system with live demonstrations at local institutions. At the same time, we must have a cost-recovery scheme available to potential users.

**Short term: 2014**

Some of the above

**Medium term: 2015**

SPARK/G (SPARK GWAS) prototype testing

**Long term: 2016 onward**

SPARK/N (SPARK Next-gen) prototype testing