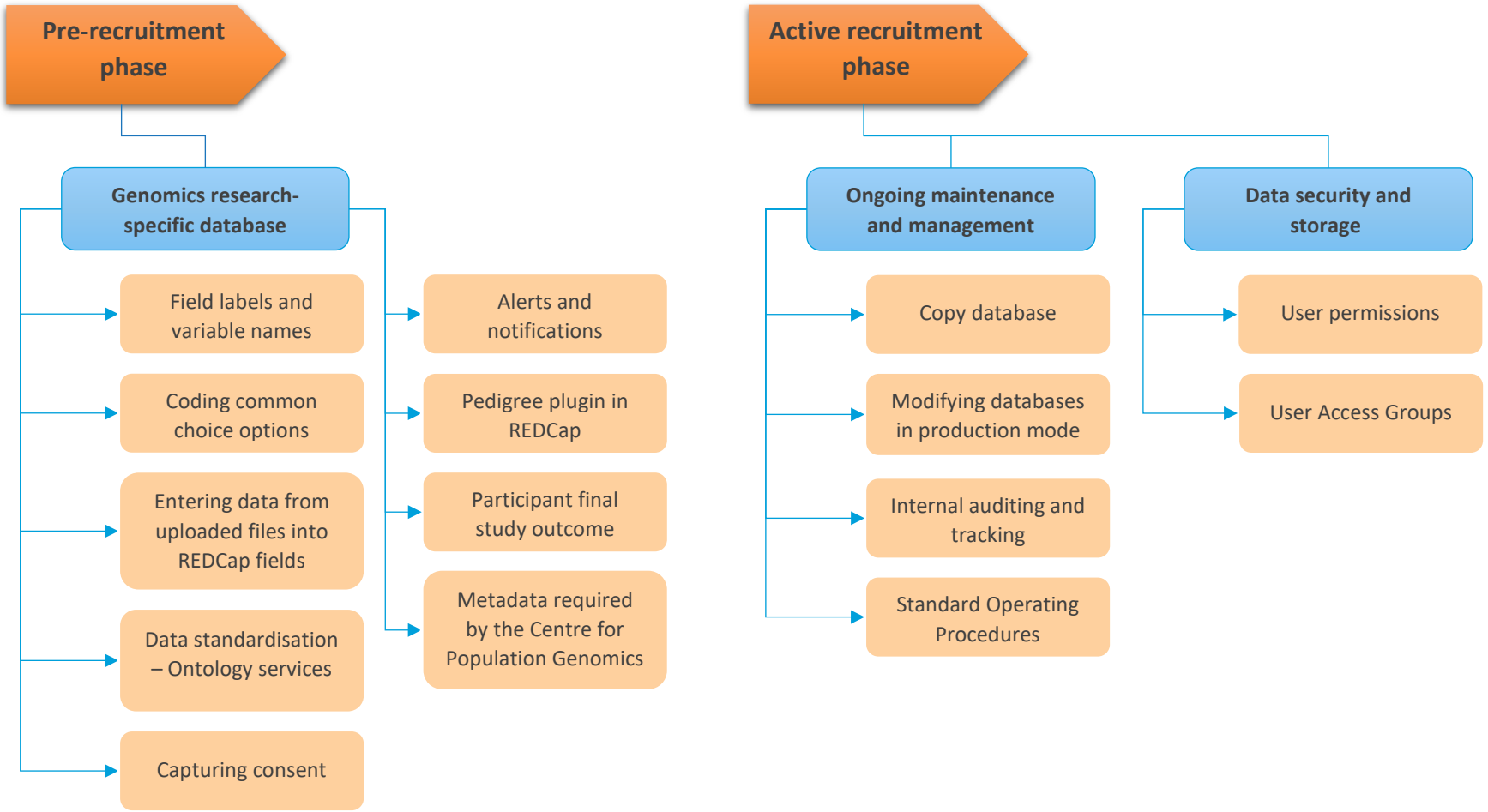
Australian Genomics study database guidelines

A resource for genomic researchers developing study databases

Australian Genomics

Database development process

(Click on a button in the flowchart to navigate to that section)



Introduction

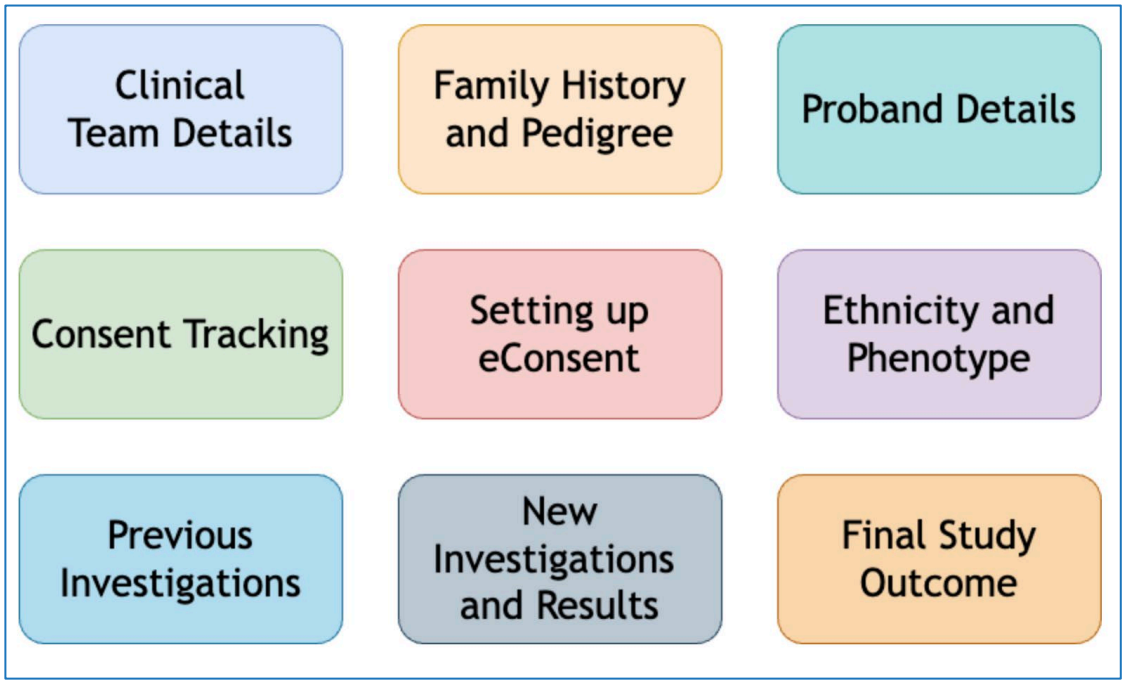
This guide provides practical guidance for genomics researchers developing study databases for their respective studies. It addresses both the pre-recruitment and post-recruitment phases of database development, with a focus on genomics-specific considerations and best practices.

Note: this guide assumes the use of the REDCap data capture platform, widely available at many universities and research institutions. However, core principles described in this document may be applied to similar platforms.

Getting started with REDCap

- A basic understanding of REDCap functionality is **highly recommended.** The REDCap consortium offers video tutorials and other resources to help researchers familiarise themselves with the platform. There are also many educational resources online, such as user manuals and question forums.
- If you require training, REDCap training may be offered by your institute or a partner organisation.
- REDCap's Shared Library is a repository of data collection instruments and forms that can be downloaded by researchers. Institutions may also have their own local repository of instruments suitable for use.

To support researchers, Australian Genomics has created a suite of templates tailored for genomics research projects. A demonstration of the participant database template can be found here.



Pre-recruitment phase – Database development

Genomics research-specific database development

Key considerations for database development

This guide expands on the key topics listed below and aims to help researchers develop a robust database tailored to the specific requirements of genomic research, minimising challenges and ensuring high-quality data for analysis.

Collaborative access

If multiple investigators will require access to the database, create a *Standard Operating Procedure* (SOP) to ensure consistent data entry and management (refer section: Ongoing maintenance and management).

Standardisation

Design your database to collect accurate, standardised data that facilitates easy analysis and reporting. Consider potential future uses of the data and design your database accordingly.

Avoid common pitfalls

Plan thoroughly before moving the database to Production (live) mode to prevent issues with real-time data collection.

Field labels and variable names

Data fields are variables in which relevant data is entered. All fields must have a unique *variable name*. Fields also generally require labels to provide human-readable context. Creating *field labels* and variable names that provide adequate context to the field is strongly recommended. Aim for consistent and informative variable names that the research team can easily interpret, which will facilitate data analysis at the end of the study. Some examples of good practice include:

- Giving variable names a unique prefix if they are in the same instrument (form), for example:
    - Fields in a participant survey instrument could have the prefix '*surv_'* (e.g., [*surv_q1_outcome*]).
- Always including field labels, even for embedded fields, to ensure that the field label appears when exporting data.
- Ensuring variable names do not have typos or formatting errors (e.g., extra spaces at the end of a sentence or paragraph).

Important: some statistical analysis packages cannot handle variable names longer than 26 characters. If you plan to analyse data using these packages, please keep this in mind and code your fields accordingly.

Coding common choice options

Within a study, many multiple-choice fields may utilise the same choice options. For choice options that appear across different fields, we recommend employing consistent code values across the database to facilitate data entry and analysis.

Common choice values should be unique enough that there is **no risk** of them appearing for other choices across the database.

For example, if there are more than 10 choices for some fields (i.e., 1-99), use a three-digit code for common answers. If there are more than 100 choices for some fields (i.e., 1-999), use a four-digit code for common answers.

Examples of common choices and a recommended choice code:

Choice label	Choice code
Yes	1
No	0
Other	999
Not applicable (N/A)	888
Unknown	777

Entering data from uploaded files into REDCap fields

In your project database, you may be uploading files important to the participant's study record including, but not limited to, test results, consent forms (if using paper or verbal consent), and request forms.

Some of the information within these files (e.g., test results, consent choices) will be important for participant tracking and reporting during the study period, so we recommend creating fields to capture the relevant information and facilitate study processes, including:

- Facilitating report creation and filtering participants (i.e., based on optional consent choices, test results, genes reported, etc.)
- Standardising study data into an exportable, machine and human-readable format for analysis and data sharing
- Searching for participant information without having to manually search individual documents.

Data standardisation – Ontology services

It is recommended that an ontology service, such as the Fast Health Interoperable Resources (FHIR) standard or BioPortal, is used to capture participants' clinical and phenotypic data in REDCap. The FHIR standard defines a set of terminology resources that provide a mechanism to access clinical terminology in a standardised way[1]. FHIR profiles contain strict terminology bindings that ensure queries can be run consistently across multiple repositories.

The FHIR standard is freely available through **Ontoserver,** a terminology server developed by CSIRO's Australian e-Health Research Centre. A REDCap plugin is available that allows databases to connect to ontologies hosted by the Ontoserver to create autocomplete-style fields in REDCap forms to capture coded data. Some of the biomedical ontologies that the Ontoserver currently contains, and are commonly used in genomic research, are:

- Human Phenotype Ontology (HPO) – **Recommended** ontology to capture phenotypic features in genomics research
- SNOMED Clinical Terms (SNOMED CT)
- Human Ancestry Ontology (HANCESTRO)
- Online Mendelian Inheritance in Man (OMIM)
- HUGO Gene Nomenclature (HGNC)

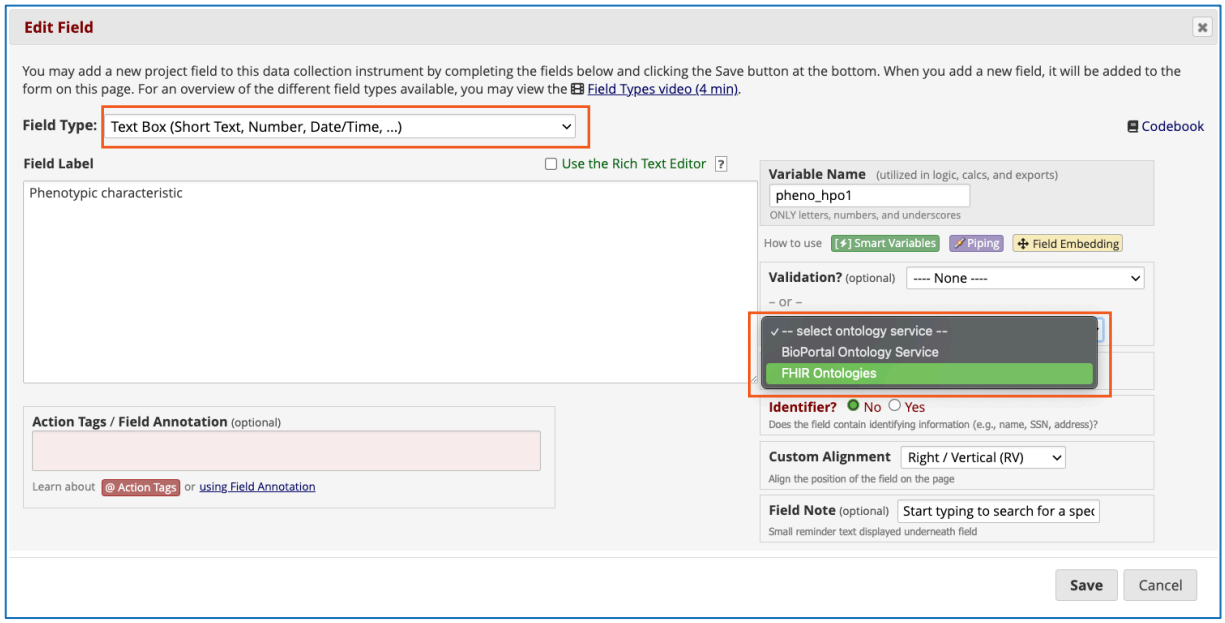
To use the FHIR Ontology External Module, it needs to be installed on your institution's REDCap server by an administrator. It is freely available to download and install via **GitHub**. You will then need to enable the module in your specific REDCap project(s). During installation, the REDCap administrator will require an ontology server to connect to. The current stable Ontoserver recommended for use is **https://tx.ontoserver.csiro.au/fhir**

If your institution's REDCap does not allow the use of External Modules, you may use the **BioPortal** field ontology, which is a native feature in all REDCap versions. To implement BioPortal ontologies, follow Step 1 below, but select 'BioPortal Ontology Service' instead of 'FHIR Ontologies' in the field validation setup. Then, choose your desired ontology from the drop-down list and save the field.

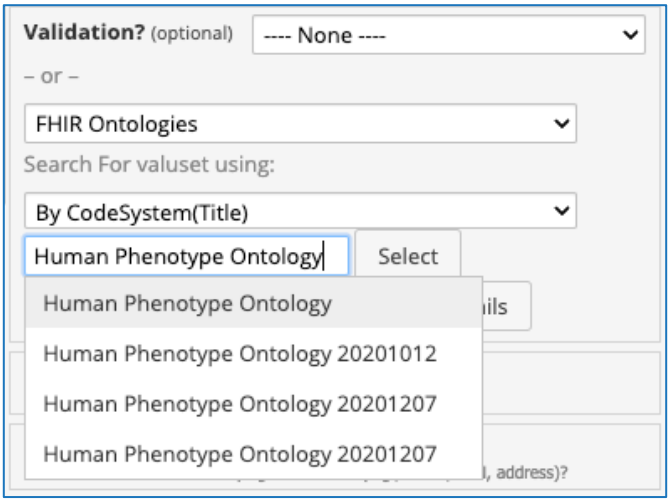
[1] Köhler Sebastian, Vasilevsky Nicole A, Engelstad Mark, Foster Erin, McMurry Julie, Aymé Ségol`ene, Baynam Gareth, Bello Susan M, Boerkoel Cornelius F, Boycott Kym M, et al. The human phenotype ontology in 2017. Nucleic acids research. 2016;45(D1):D865–D876.

How to create a FHIR autocomplete field in REDCap

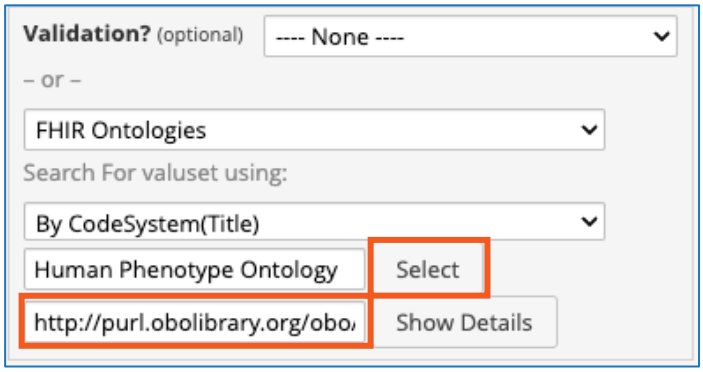
1. Create a text box field, and in the '**Validation**' section, select the **FHIR Ontologies** validation.



2. Below this, in the dropdown bar '**Search For valuset using:**' select '**By CodeSystem (Title)**'. In the field below that, search for the preferred ontology (e.g., Human Phenotype Ontology)



3. Select the resulting ontology and click on the '**Select**' button. A link should appear in the field below.



4. Save the field.

You can create as many fields as required to capture the number of clinical characteristics needed. However, some journals now limit the number of ontological terms per participant due to data privacy. In this case, we recommend adding 3-5 primary phenotype capture fields that describe the patient's overall phenotype that can be used in publications. If more phenotypic features are required for genomic analysis, these may be captured but as secondary fields for study use only.

Important: The ontologies listed are broad value sets that include a variety of terms, some of which may not be relevant to your project. It is possible to exclude values that are not relevant to your study by implementing a more restrictive HPO value set. To organise this, please contact the Ontoserver administrators at ontoserver-support@csiro.au. To view all available value sets in the TX Ontoserver, please visit the CSIRO Shrimp server.

References to cite when using the module

When using the REDCap FHIR Ontology External Module, please use the following acknowledgement text for any publications:

Some of the data in this study was collected using the REDCap FHIR Ontology External Module, developed by CSIRO and Australian Genomics, and Ontoserver, CSIRO's FHIR terminology server.

Additionally, cite these papers in publications (both describe technologies used by the plugin):

- *Metke-Jimenez, A., Steel, J., Hansen, D. et al. Ontoserver: a syndicated terminology server. J Biomed Semant **9**, 24 (2018). https://doi.org/10.1186/s13326-018-0191-z*
- *Metke-Jimenez A, Lawley M, Hansen D. FHIR OWL: Transforming OWL ontologies into FHIR terminology resources. AMIA Annu Symp Proc. 2020 Mar 4;2019:664-672. PMID: 32308861; PMCID: PMC7153051.*

Capturing consent

Capturing paper and verbal consent forms

A member of the study team can upload a scanned copy of the completed consent form(s) using a '**File upload**' field. Before the consent form is uploaded, it is important to ensure that the participant's information in the form aligns with their record.

We recommend including additional fields to capture key information from the uploaded consent form that may be required for reporting and analysis, including (but not limited to):

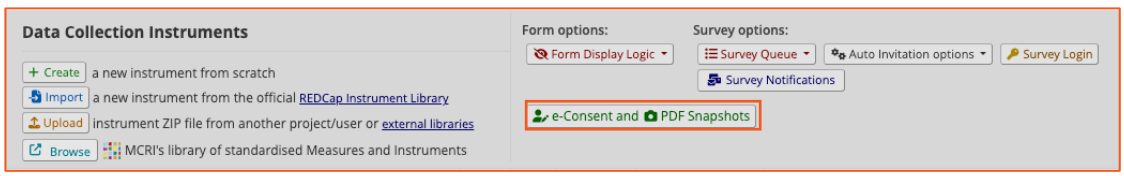
- Date consent obtained
- Optional consent choices
- Who signed the consent
- Consent form version number

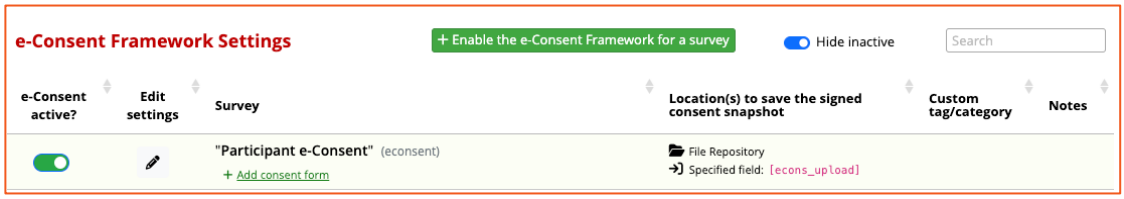
e-Consent using the 'Auto-Archiver + e-consent framework in REDCap

Electronic Consent (*e-Consent*) is used to consent participants using a computer-based consent form rather than traditional paper documentation. Consent forms can be developed as a REDCap survey and completed via computer, mobile phone, or tablet.

How to set up e-consent in REDCap

1. Create a new instrument in your REDCap project and enable it as a survey.
2. Create all fields that are required to match the consent form approved by your ethics committee.
    a. It is **highly recommended** creating a hidden date field using the @HIDDEN-SURVEY action tag at the end of the e-Consent survey to capture the date of consent.
3. Set up a signature method in the survey. Participants can 'sign' their consents using several different methods (this field **must** be a required field):
    a. Typing in their full name and submitting the form.
    b. Utilising REDCap's '**Signature**' field type (i.e., digital signature).
    c. Using a PIN number assigned to them by the study team.
4. Set up the survey settings as appropriate for your project.
    a. **Note:** If you wish to allow participants to save and return to the survey later, requiring a return code is **strongly recommended**, as the e-Consent survey will contain identifier fields.
5. In REDCap's *Online Designer*, click the '**e-Consent and PDF Snapshots**' button, and follow the prompts to add and edit e-Consent settings for individual instruments:





Important: The signature process will **NOT** be implemented by REDCap automatically, so it is your responsibility as a survey administrator to construct your survey using one of the methods above for the signature to get captured appropriately.

Consent via the CTRL platform

The Australian Genomics-designed web application **Control (CTRL)** is an online dynamic consent platform that encourages ongoing participant engagement with the study. Instead of signing a once-off consent form, CTRL permits participants to review and update their decisions about research participation in real time and enables more granular decision-making about the future use of their research data.

While CTRL is a standalone web app, it can be linked to REDCap databases via an API. This enables participant consent choices to be updated in real-time within their REDCap records.

CTRL is hosted by the Garvan Institute of Medical Research and is available for uptake from GitHub. If you would like to enquire about how to implement CTRL to your project, please contact the CTRL development team at ctrl@garvan.org.au.

For more information and to access a demonstration version of CTRL, please visit the Australian Genomics website.

Alerts and notifications

Overview

Alerts and notifications allow project teams to send customised email notifications to the desired recipient(s). Alerts can be set up to be triggered or scheduled when an instrument is saved and/or based on conditional logic whenever data is saved or imported.

In REDCap, the *Alerts and Notifications* feature allows for greater complexity compared to *Automated Survey Invitations* and has more capabilities. They can apply to both data entry forms and surveys, and they also allow for more options regarding who can be the recipient of a notification (e.g., project users, survey participants, etc.).

Within the Australian Genomics template databases, some common alerts have been set up to assist with database development. These include:

- *Recruitment database*
    - New participant submitted – alerts study team to review a new submission that has been created.
    - Participant accepted – alerts the referring clinician that their patient was accepted into the study.
    - Participant declined – alerts the referring clinician that their patient was not accepted into the study.
- *Participant database*
    - New participant added – alerts study team that participant data from recruitment database has been copied to the participant database.
    - e-Consent invitations – alerts patient and/or additional relatives to complete the relevant e-Consent form for the study.

Important considerations when using alerts

- It is recommended to always use a test database with dummy records when testing new alerts to avoid sending unsolicited emails to participants.
- If an alert is created part way through the study and you would like it to apply to prospective participants only, you will need to create a condition in the logic (**Step 1: triggering the alert**) that blocks the alert for retrospective participants who shouldn't receive it. Alerts are very sensitive so they should be tested thoroughly before being implemented.
- Consider setting up alerts to be sent with a time delay (e.g., 30 minutes) so that you have some time to visit the '*Notification log*' and check that it has been set up properly.
    - **Note**: Time delays will **NOT** work if the alert conditions include a date as a trigger field and this date has already passed the duration of the set time delay. For example, if the alert has been set to send 30 minutes after the date entered, and the date entered was two days before the current date, the alert will send immediately, as it has been more than 30 minutes since the date entered.

Pedigree plugin in REDCap

The Pedigree Editor External Module converts a '**Notes**' field into a pedigree editor using the Open Pedigree Tool by PhenoTips. The aim of the pedigree tool is to facilitate the generation of machine-readable pedigrees (as opposed to hand-written pedigrees) and enhance interoperability between REDCap and other software tools (e.g., Electronic Medical Records) or risk analysis tools (e.g., BOADICEA breast cancer risk prediction tool).

Note: your REDCap administrators will need to install this module before being active in your project. If your institution's REDCap does not allow external modules, you will not be able to use the pedigree editor tool.

The edited pedigree can be viewed within REDCap and exported as any of the following formats:

- PED
- GA4GH FHIR
- SVG
- PDF
- DADA2

Within the REDCap field itself, the diagram will be saved as a FHIR Composition JSON code, which can be useful for importing into compatible programs, such as across healthcare systems and other software tools.

Setting up a pedigree field

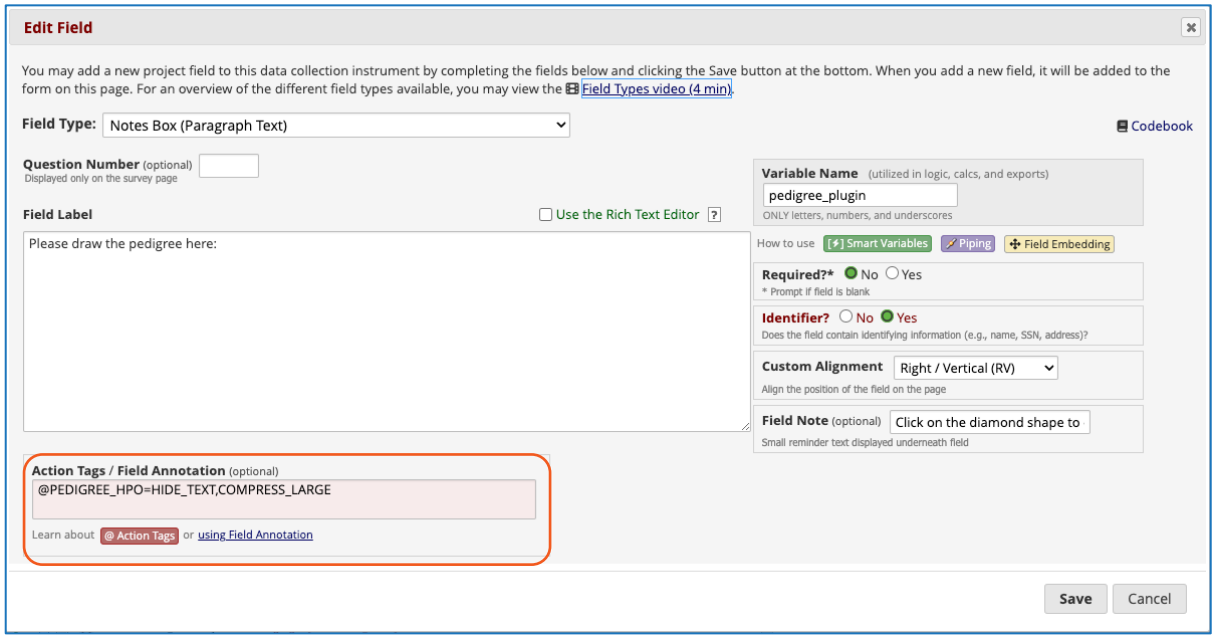
To use the pedigree editor, create a **Notes Box** field and add one of following three action tags:

- @PEDIGREE_HPO - Marks a field to be a pedigree editor using the HPO and OMIM coding systems for phenotypes and disorders.
- @PEDIGREE_SCT - Marks a field to be a pedigree editor using the SNOMED-CT coding system for phenotypes and disorders **(recommended)**.
- @PEDIGREE – Marks a field to be a pedigree editor being agnostic of the code system, i.e. it will allow HPO / OMIM and SCT coding for phenotypes and disorders in separate fields within the editor.

The default '*Hide Text*' and '*Compress Data*' options can be overridden in the action tag by appending '=' plus a comma-separated list of options.

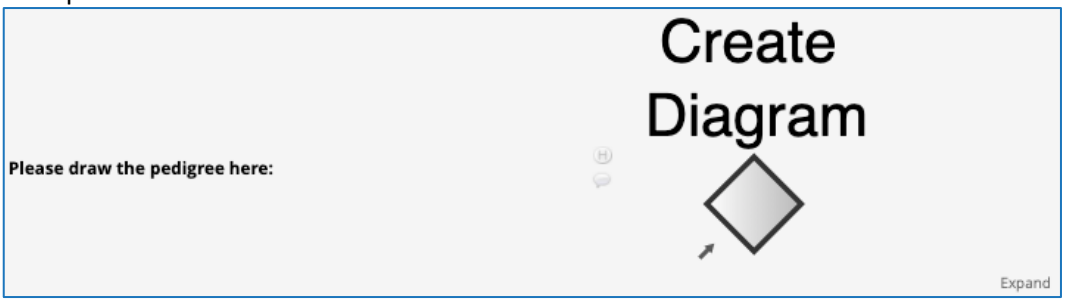
- HIDE_TEXT - Hide the text area.
- SHOW_TEXT - Show the text area.
- NEVER_COMPRESS - Never compress the data stored in this field.
- COMPRESS_LARGE - Compress the data if it exceeds 65K characters.
- ALWAYS_COMPRESS - Always compress the data for this field.

Example: an action tag of @PEDIGREE_HPO=HIDE_TEXT,NEVER_COMPRESS will use the HPO code system and hide the text area for the field and prevent compression of the data stored in the field.

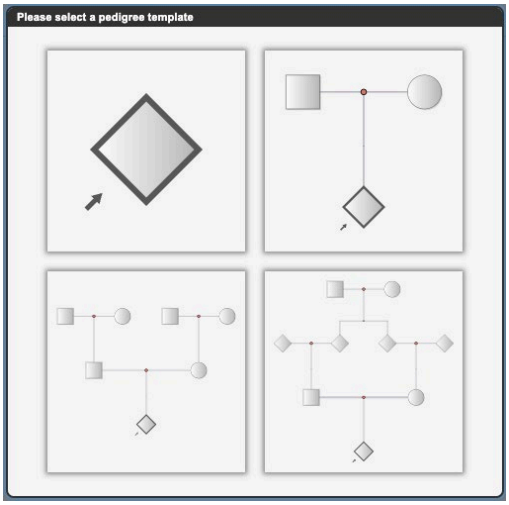


How to use the pedigree tool

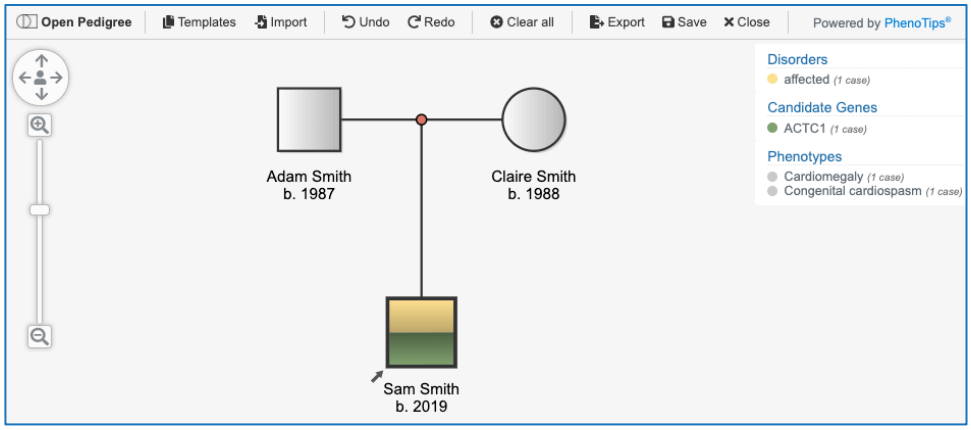
1. In a participant's record, find the pedigree field and click '**Create Diagram**'. The pedigree tool will open in a new tab.

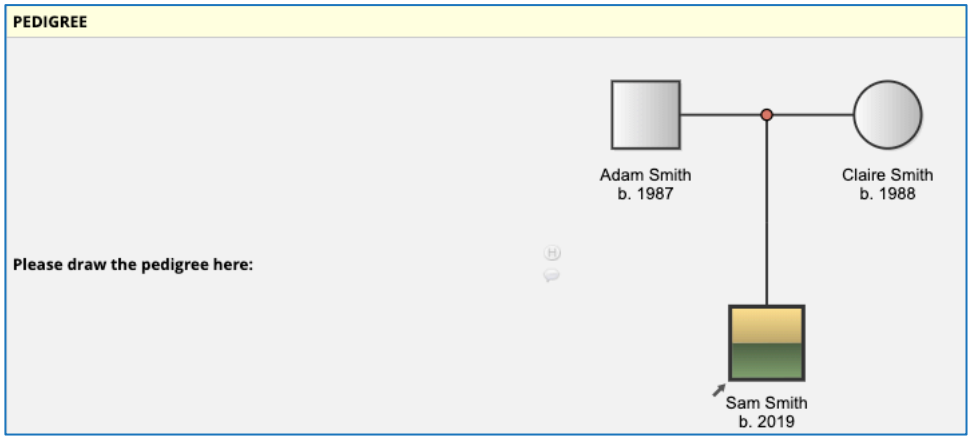


2. When you first open the tool, you will be able to choose a starting template based on the type of information you wish to capture. Once a template is selected, you may begin editing the pedigree.



3. You may click on any family member to add names, date of birth, living status, sex, phenotypic features and/or diagnoses, carrier status, and other comments. You can also click on lines protruding from each family member to expand the pedigree (e.g., adding siblings, parents, or children).





4. When you have completed the pedigree, or wish to save your progress, click the '**Save**' button and it will close the tab and return to the participant's record in REDCap. If the pedigree has saved correctly, you should see an image of the pedigree in the field.
5. If you wish to edit the pedigree, click on the pedigree image in the REDCap field and it will take you back to the editor.
6. To export the pedigree as an image, you must do this on the pedigree editor page. Click '**Export**' and choose an appropriate format (for a human-readable image, we recommend a PDF).

Participant final study outcome

For studies that are conducting multiple tests for participants, there may be challenges with collating all test data into a report for analysis. In this case, it is recommended to utilise an instrument that includes a summary of findings across each participant test.

This instrument may contain the following:

- A checkbox field listing all available tests in the study
- Final participant result (and, if positive, the test that found and/or confirmed the result)
- Final result disclosure and case resolution notes.

An example of a findings summary instrument can be found in the Australian Genomics Template Database Demo.

Metadata required by the Centre for Population Genomics *seqr* platform

For projects that intend on storing their genomic data on the Centre for Population Genomics' (CPG) ***seqr*** platform, there is a minimum set of required metadata fields that will need to be captured in REDCap to ingest genomic data and link it to the participants' demographic and phenotypic data.

The full and current metadata requirements for CPG can be found on GitHub here.

If you are building a REDCap database to use in coordination with CPG *seqr,* we suggest you contact CPG to discuss the details of their project.

Contacts

Cas Simons, Rare Disease Lead

Email: cas.simons@populationgenomics.org.au

Rocio Rius, Rare Disease Team

Email: rocio.rius@populationgenomics.org.au

Active recruitment phase

Once a study is in the active recruitment phase and recording real data, project owners are responsible for overseeing appropriate data management and security, and auditing data entry by users.

Our recommendations for best practice when capturing live data in a project, with special emphasis on the use of a study REDCap database, are given below.

Ongoing maintenance and management

Copy database

A copy database is an identical copy of your real database, without participant data. Copying a REDCap database can be performed under the '*Other Functionality*' tab from the project home page.

The copy database allows project owners to test new instruments or modifications without affecting the data in the live (production) database. This is particularly useful when testing alerts and notifications, mitigating the risk of accidentally sending alerts to existing participants. Once testing is complete, the new modifications can be implemented in the real database.

Another benefit of copying a database is to retain a backup prior to the database being moved into Production and real data being entered. Copy databases can also serve as instrument templates for future projects.

Modifying databases in production mode

It is possible to make database modifications once a database is in *Production* mode and users are entering real data, however caution and thorough testing is recommended.

We recommend that your database is fully developed, tested, and is functioning appropriately before moving the database into *Production* mode. This minimises any errors in the database prior to entry of real data and prevents important existing data from being lost. Changes that may affect existing data will require administrator approval before being implemented.

Examples of database modifications that will require administrator approval are:

- **Deleting fields**: deleting fields will also delete any previously recorded data. An alternative is to use an Action Tag (e.g., @HIDDEN) to hide the question if you know that data has been recorded for that field.
- **Deleting/modifying existing choice options**: deleting or modifying existing choice options within a field will affect all the data associated with that choice option. If a particular choice option is no longer valid, you may use an action tag @HIDECHOICE to hide the unwanted option.
    - Avoid changing the numbers associated with common choice options: this will associate previously stored answers with a new (and different) answer option.

Internal auditing and tracking

During the active recruitment phase, study personnel can use their REDCap database to track the completion of project milestones and provide updates about recruitment and data collection to project teams.

The REDCap database provides a central location where study coordinators can track participant consent, withdrawals, changes to personal and sensitive information, instrument completion and changes to their clinical team, etc.

Educating project teams on how to use the study's participant database (via a *Standard Operating Procedure*), or who to contact to enter data into the database, aims to protect participant information and reduces the likelihood of team members inappropriately storing participant information outside the database (e.g., on local computers in unaudited spreadsheets).

Depending on the size of your project, you may need multiple people tracking incoming data. Individual projects will need unique ways for tracking data and depends primarily on the data source (e.g. electronic or paper forms, web-based platforms).

Methods to track data within a REDCap database include:

- Generating reports using the **'Data Exports, Reports, and Stats'** tool.
- Creating **'Alerts & Notifications'** sent to the project team at data collection timepoints or project milestones.
- '**Data Logging'** tool to audit who has accessed and modified participant data.
- External, standardised spreadsheets or reports using input from REDCap reports created using the **'Data Exports, Reports, and Stats'** tool.

Best practices for auditing and tracking data

- Regularly track and audit data collection throughout active recruitment (do not wait until the end of data collection to begin auditing data).
- If uploading data initially captured external to REDCap (e.g., paper consent forms), review data for accuracy and completeness before adding to REDCap.
    - **Note:** if collecting data using paper forms, it is **strongly recommended** that this data be input into specific REDCap fields to ensure standardisation across the database and easier data analysis.
- Include reliable checks/alerts to ensure collected data is complete, and data is not missing or entered incorrectly.

Standard Operating Procedures

A *Standard Operating Procedure* (SOP) is a document informing the project team on how procedures will be implemented throughout the project duration. The SOP document should contain detailed instructions on how to manage data entry and collection of real data into the database, and should reflect the processes outlined in the study's Protocol.

An SOP document should be created during the *Database Development* phase, and prior to a database being moved into Production. SOPs for projects that utilise REDCap databases will generally include:

- Scope and purpose of the document
- Workflow diagrams or processes
- Terminology and abbreviations
- Procedures and associated steps for entering real data in instruments
- Responsibilities of various project team members
- File names, links to policies, videos or tutorials (if available)
- Communication methods
- Database screenshots (if applicable)

Data security and storage

User permissions

To control access to Personal Identifiable Information (PII), users can be limited in accessing certain data features, such as editing or exporting data, or deleting records. This is achieved by assigning user rights to individuals, or in REDCap by creating '*User Roles*' that group specific limitations and access rights, and then assigning individual users to that '*User Role*' group.

Best practice is to set user rights according to the minimum requirements required by that user to access and use the database. This mitigates the risk of data being accessed inappropriately and controls the possibilities of data accidentally being lost or compromised. Only project owners should have the highest access permissions in the database. Other study personnel (e.g., clinicians, genetic counsellors, lab personnel, etc.) should only be able to access instruments and data relevant to their role in the study.

User Access Groups

Access Groups are another useful tool to limit access to project data to certain user groups (e.g., project owners, data entry personnel, clinicians, etc.). Access groups can also be helpful for multi-site studies, where users only require access to data collected at specific sites.

In REDCap, these are known as *Data Access Groups (DAGs)*. Different DAGs can be assigned to users within the same 'User Role' group, even if the users within the group are from multiple institutions, allowing further customisation of access to records. For example, clinicians involved in a study would have user rights associated with a 'Clinicians' user role. However, a study may have clinicians across different study sites. Therefore, assigning DAGs to each clinician within this user role prevents clinicians accessing records at other sites.

A user can also be assigned to multiple DAGs (e.g. if a clinician works at multiple study sites). This is done by using the *DAG Switcher Feature* to switch between DAGs as needed.