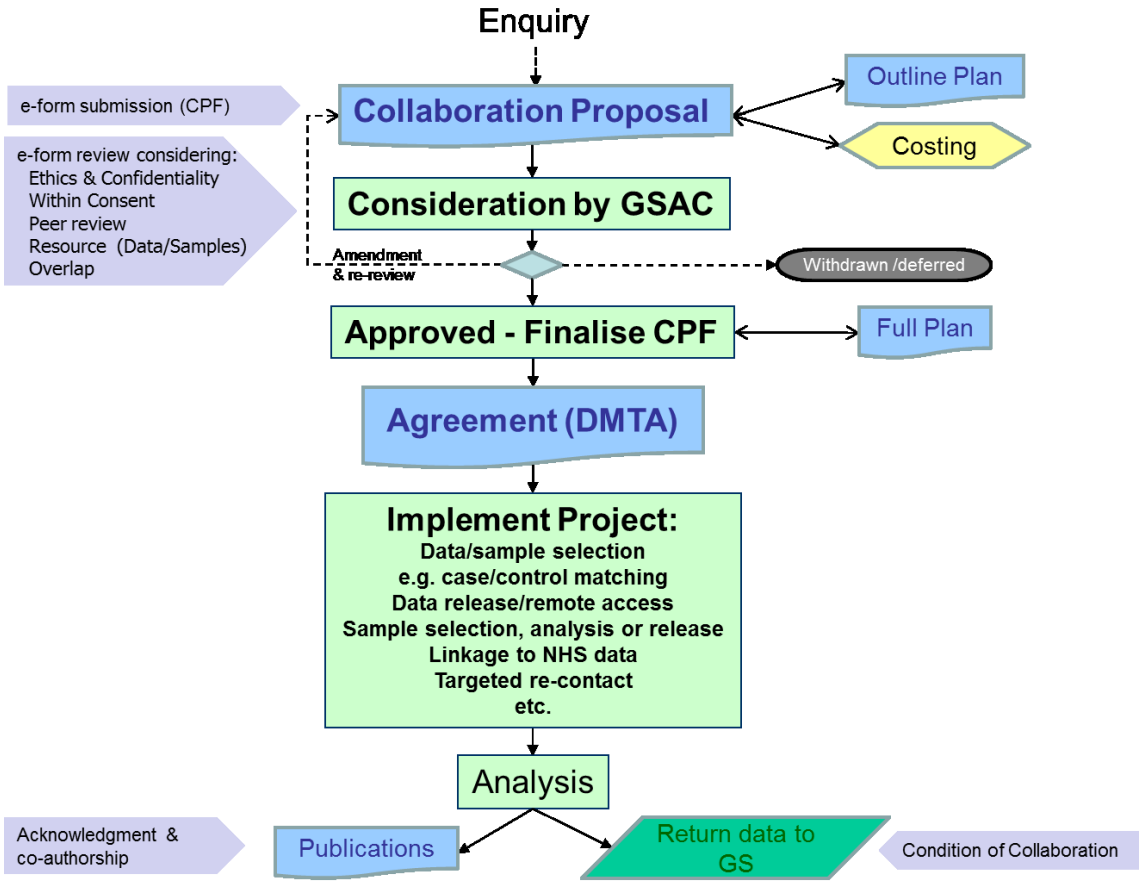
Generation Scotland – provision of access to GS resources for collaborative research projects

GENERATION SCOTLAND (GS) is a multi-institution, cross-disciplinary collaboration (bound by a Collaboration Agreement) between the Scottish University Medical Schools (Edinburgh, Glasgow, Aberdeen and Dundee), The NHS Health Boards of Grampian, Lothian, Greater Glasgow and Tayside, and the Information Services Division of NHS National Services Scotland.

GS has created a resource of data and biological material collected from over 30,000 participants within three studies: Scottish Family Health Study (GS:SFHS), Genetic Health in the 21st Century (GS:21CGH) and the Donor DNA Databank (GS:3D). The resource is made available to researchers through managed access. Research Tissue Bank approval has been obtained for each of these collections to provide generic ethical approval for a wide range of uses within medical research and a robust governance process established and implemented by the GS Access Committee.

Approvals

All proposals for use of the resource require completion of a "Generation Scotland Collaboration Proposal Form" (online form, access available on request) which is reviewed by the GS Access Committee. The principal function of this Committee is to ensure good governance and economic and appropriate use of the GS resources. Proposals requesting linkage to other datasets require additional approval e.g. PBPP approval for linkage to NHS datasets. Having obtained approval from the Access Committee a signed "Data / Material Transfer Agreement" and proof of any other approvals (ethics, PBPP etc.) is required before release of data and/or samples to the recipient. This process is summarised below:



Study Data

GS:SFHS

A unique study ID was applied to the study data and samples at the point of collection. Personal information for study participants (including the Health Index number, CHI) is held separately from all other study data along with an encrypted version of the CHI. The key to the encrypted linkage is held in the NHS system and is not accessible by GS or collaborating researchers. Personal data is held separately from study data, and samples collected in the GS:SFHS study are also held separately within the institutions involved in its collection and processing.

- Personal information including CHI and encrypted version of CHI - held within the NHS N3 system at the Health Informatics Centre (HIC), University of Dundee
- Phenotype and genotype data – held on the University of Edinburgh systems at IGMM (Institute of Genetics and Molecular Medicine
- Sample management data – held at the Wellcome Trust Clinical Research Facility, University of Edinburgh, using a Laboratory Information Management System (LIMS).
- NHS medical records are held by ISD (Information Services Division under the governance of the PBPP (Public Benefit and Privacy Committee

Linking of datasets

After an access request is approved, the Data Sources work together to provide a linked anonymised subset of data (excluding personal information) directly to the named Recipient for analysis as part of the project.

The data from each Data Source is indexed by a unique ID and the Recipient is provided with a file that links these unique IDs. Using an application provided by GS ("GSlink") a set of tables using a single set of identifiers can be created by the Recipient.

An independent Linker at HIC holds information on identifiers that link the various data sources and all study data is held separately.

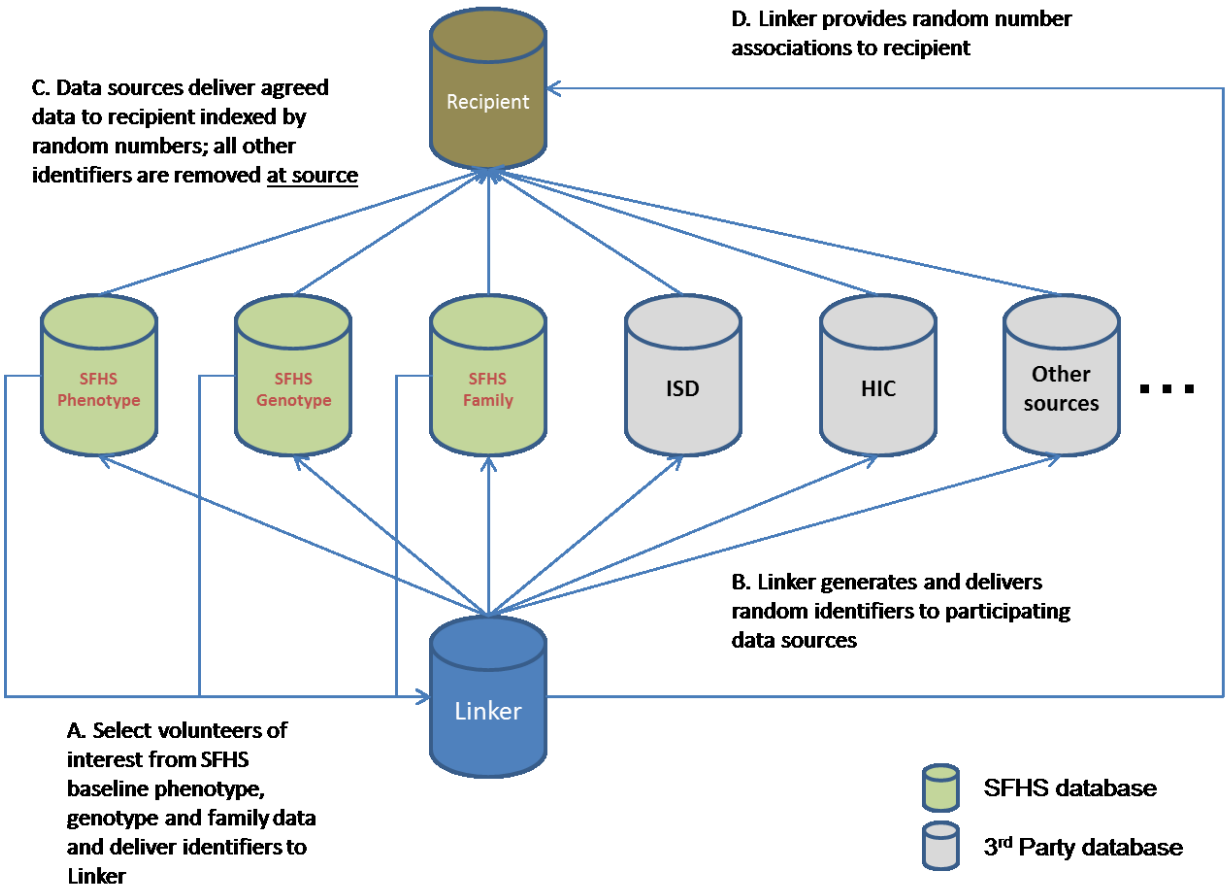
The steps involved are set out below.

A. Selection of study data records and fields to be included in the release is based on criteria set within the approved Collaboration Proposal Form. GS identifiers for the required records are passed to the Linker by the GS data analyst.

B. The Linker generates a separate set of random identifiers for each set of records to be released from each Data Source and sends this, along with the corresponding data source identifiers, to each Data Source involved in the linkage.

C. Data Sources prepare the required data indexed by the new random identifier and send this directly to the Recipient. All other identifiers are removed.

D. The Linker sends a password protected file linking the random identifiers used by each data source to the Recipient, to enable them to link the datasets together using GSlink. The newly created linked dataset is only available to the Recipient.



The mechanism used to achieve this works for any number of GS databases and datasets from external parties are easily incorporated. The only requirement is that the relationship between identifiers is known within the collection of databases and an appropriate identifier is available for each database (e.g. CHI for linkage with NHS records, ACONF ID for linkage with ACONF datasets, etc.). Subsets of data from external sources indexed by the CHI are selected based on CHI but brought together with study data and indexed by a new unique ID to maintain anonymity. The Linker provides a list of encrypted CHI and a new set of identifiers to the NHS at HIC. Using the encryption key a list of the CHI and the new set of identifiers is then sent to the external source to allow selection of the required records and application of the new identifiers before release to the recipient. The CHI will not be included in data released to the recipient.

Key features:

- A new set of random IDs are applied to each new release, ensuring data released for different projects cannot be linked.
- This mechanism allows each Data Source to maintain control over the use of its datasets. To minimise risk, information can be provided on the data fields from other datasets involved in the linkage to each Data Source if required.
- Datasets from various sources/external parties can be linked.