Department of Twin Research

Data Management

Our data management processes involve the collection, administration, technical handling, storage and dissemination of data and information.

Our data management practices include:

- Clearly defined roles and responsibilities for those associated with the data, in particular of the data owner, data requestors, data providers and custodians.
- Data quality procedures (e.g., quality assurance, quality control) at all stages of the data management process.
- Verification and validation of accuracy of the data.
- Documentation of specific and descriptive metadata for each dataset.
- Adherence to all ethical issues regarding data management stipulated by the department.
- Defined procedures for updates to the information system infrastructure (hardware, software, file formats, storage media), data storage and backup methods and the data itself.
- Ongoing data audit to monitor the use and assess effectiveness of management practices and the integrity of existing data.
- Data storage and archiving plan and testing of this plan (disaster recovery).
- Evolving data security approach of layered controls for reducing risks to data including audited access logs.
- Clear statements of criteria for data access.
- Documented delivery procedures for data that is available and useable to users including Open Access where applicable

Activities include:

- Policy and Administration
  - data policy
  - roles and responsibilities
    - data owner
    - Approval Committee
    - data custodian
    - data provider
    - data requestor
- Collection and Capture
  - data quality
  - data documentation and organisation
    - dataset titles and file names
    - file contents
    - metadata
  - data standards
  - data life-cycle control
    - data specification and modelling (database design)
    - database maintenance

        - data audit
        - data storage and archiving
- Longevity and Use
    - data security
    - data access, data sharing, and dissemination
    - data publishing

Policy and Administration

Data Access Policy

Our current data access policy document can be found at:
http://www.twinsuk.ac.uk/wp-content/uploads/2012/05/TRU_Policy_Collaborations.pdf

Roles and Responsibilities

The DTR endeavours to:

- define clear roles associated with each data management function
- establish data responsibility throughout all phases of any data project
- ensure data accountability
- ensure data quality, integrity and security is maintained at all times by the relevant personnel

Data Owner

Over the past 23 years, we have collected extensive biological and questionnaire data on several hundred phenotypes related to common diseases or intermediate traits. The cohort, made up of over 12,000 twins, is also the most highly and deeply genotyped twin resource in the world. Ownership and intellectual property rights belong to the DTR, King's College London. We will consider dividing intellectual property rights where applicants will be making a particular contribution. Any such division must be considered by the Approval Committee and agreed before the project starts.

Approval Committee

This committee represents the DTR and meets weekly to consider submission of data request proposals and papers. Before data is released to the requestor the conditions of use as set out in a signed memorandum of agreement/application form must be agreed to.

Data Custodians

Specific roles exist within this realm:

- IT Manager
- Data Manager
- IT Support
- Database Administrator
- Database/Application Developer

Data custodians ensure that important datasets are developed, received from external sources, maintained and are accessible by the appropriate users. The IT and Data Manager are in charge of overseeing all the aspects of data management and help to ensure that datasets maintain their integrity and do not become compromised.

IT support ensures the maintenance and security of hardware, software and network equipment used to access the data whilst the Database Administrator ensures the safety, integrity and security of the data and database management software.
The Database Developer creates applications that help extract data, provide metadata and generally turn datasets into useful information. They also provide the framework for the security of our data and an infrastructure to realise our Open Access ethos.

Data Provider

If datasets are required that are not part of the Open Access facility or of a complex nature, the Data Provider will analyse the needs of the Data Requestor and provide a tailored dataset. They will then provide support to the requestor to bridge communications between them and the Department.

Data Requestor

Data defined as Open Access can be accessed via a portal on our website. Researchers can register to be added to the system and having received their login name and password can access the data as and when required. Access is logged but not restricted.
Researchers needing access to other data from the DTR or who are interested in proposing a collaboration with us, complete a Data Access Application Form found on the website. Background information can be found in the DTR Policy Document for data access. The proposal should specify the data required clearly. Individual variables need to be listed with an appropriate justification describing the aims/hypothesis of the project for which the data are requested.

Collection and Capture

Data Quality

Data quality is affected by the way data is entered, stored and managed. It is defined as fitness for use and to serve its purpose.

The “Data Custodians” at the DTR (See above) are tasked with maintaining our data quality at all times. This requires inspecting the data on an on-going basis with periodic audits, “cleaning” and “harmonising” it. This may involve updating, standardisation, and removing duplicate records. A single view of the data is created if it is stored in multiple disparate systems. Quality as applied to data has been defined as — fitness for use or — potential use (i.e. fit for purpose).

Loss of data quality can occur at any stage in the data management process, reducing its applicability and use:

- Data capture and recording at the time of gathering
- Data manipulation prior to digitisation (smudging of answers on a questionnaire, loss of page, etc.)
- Digitisation of the data
- Documentation of the data (capturing and recording the metadata)
- Data storage and archiving
- Cleaning & Harmonisation of the data
- Data presentation and dissemination
- Using the data (analysis, integration, etc.)

Our data quality is measured against the following:

- Accuracy
- Precision
- Reliability
- Repeatability
- Reproducibility
- Currency
- Relevance
- Completeness
- Ability to audit

Quality control (QC): An assessment of quality based on internal standards, processes, and procedures established to control and monitor quality.

Quality assurance (QA): An assessment of quality based on standards external to the process and involves reviewing of the activities and quality control processes to insure final products meet predetermined standards of quality.

The DTR data team carry out both QA and QC on a continuous basis. QA procedures maintain quality throughout all stages of our data development whilst QC procedures monitor and evaluate all data products that are created by departmental initiatives.

Although a data set containing no errors would be ideal, we endeavour to reach 85%-90% accuracy considering 2 factors: frequency of incorrect data fields or records and significance of the data field with the error.

We can detect errors easier if the dataset expectations are clearly documented and what constitutes a significant error is understood. For example, a misplaced decimal point in 999 leading to 99.9 is much more significant than an incorrect decimal value in 999.99 leading to 999.98. So a pragmatic approach to data quality standards has to be taken. However, in certain circumstances one incorrect digit in a six-digit number can lead to a whole different selection of items! The significance of an error can vary both among datasets and within a single dataset.

There are two fundamental types of errors in a dataset.

- ***Errors of commission*** include those caused by data entry or transcription, or by malfunctioning equipment. These are common, fairly easy to identify, and can be effectively reduced up front with appropriate QA mechanisms built into the data acquisition process, as well as QC procedures applied after the data has been acquired. It is vital to employ as many checks as possible to reduce these “easy to identify” errors.
- ***Errors of omission*** often include insufficient documentation of legitimate data values, which could affect the interpretation of those values. These errors may be harder to detect and correct, but many of these errors should be revealed by rigorous QC procedures. Here it is vital to employ expert knowledge and apply their valid information when defining the data with metadata. For example, at the DTR the Principal Investigator of any project that gathers data is engaged at the beginning to describe all data and the equipment that collects the data including units and ranges.

As part of the QC process, data quality is assessed by applying verification and validation procedures. Both verification and validation are important components of data management that help ensure data is valid and reliable.

- ***Data verification*** is the process of evaluating the completeness, correctness, and compliance of a dataset with required procedures to ensure that the data is what it purports to be.
- ***Data validation*** follows data verification, and it involves evaluating verified data to determine if data quality goals have been achieved and the reasons for any deviations.

Our data entry officers (usually outsourced to professional data entry companies) perform data verification checks and make sure that the digitised data entered matches the source data. Data is often automatically scanned but verified by a human being. They do not necessarily need to be familiar with the actual information the data purports to. A “double blind entry” method can also be employed to reduce these errors down to almost 0%. The more important step in the process is validation checks performed by more expert staff making sure that the data makes sense. For example checking against the units and ranges provided by the PI. This step can also contain manual and automated algorithmic batch checks.

Principles of data quality are vital to the data team of the DTR and are applied at all stages of the data management process (capture, digitisation, storage, analysis, presentation and usage). We apply two key concepts to the improvement of our data quality: ***Prevention and Correction***. Error prevention is closely related to both the collection of the data and the entry of the data into the Phenobase. A well designed data collection system, through standardised survey methodologies (both paper and online) will help achieve this. However, no matter how vigilant our team is using prevention techniques, the fact remains that errors in large datasets will continue to exist and data validation and correction/cleaning has to be performed.

Documentation is the key to good data quality. Without good documentation, it is difficult for users to find what they are looking for and to determine the fitness for use of the data. It is also difficult for custodians to know what and by whom data quality checks have been carried out and audit their data.

Before any data is allowed to be imported in the Phenobase we require clear documentation to be provided and stored with the data. This requires (but is not limited to) two parts without which data quality would be compromised. The first is the metadata that records information at the dataset level describing what the data represents. The second is tied to each record and records what data checks have been done and what changes have been made. This is vital to the audit processes and for harmonisation with similar data collected in the future.

Data Documentation

Data documentation is critical for ensuring that datasets are useable well into the future. Left long enough data becomes almost unusable without comprehensive definition and description. All datasets should be identified and documented to facilitate their subsequent identification, proper management and effective use and to harmonise with future collections of similar data.

The objectives of data documentation are to:

1. Ensure the longevity of data and their re-use for multiple purposes
2. Ensure that data users understand the content, context, and limitations of datasets
3. Facilitate the discovery of datasets
4. Facilitate the interoperability of datasets and data exchange.
5. Ensure harmonisation with future collections in longitudinal studies

One of the first steps in our data management process involves entering data or converting data into an electronic format. The following data documentation practices are followed to facilitate the retrieval and interpretation of datasets.

Naming rules and short descriptions

Dataset titles and the corresponding file names will be descriptive as these datasets may be accessed many years in the future by people who will be unaware of the details of the project or study. The short description included with the file will include enough information to uniquely identify the data file. The short description will contain information such as project acronym or name, study title, location, Principal Investigator, start and end dates of study, data type, version number, and file type. As file names cannot usually be overly long, a short description is required and will facilitate the ease of search. Including a data file creation date and version number enables users to quickly determine which data they are using if an update to the dataset is released.

Contents

In order for others to use our data they must understand the contents of the dataset including the phenotype names/labels, units of measure, formats, error codes and definitions of coded values. All these are included in Phenobase and recorded with the data/datasets themselves. All information recorded for phenotypes within Phenobase will include a phenotype code, a phenotype name and where appropriate a phenotype unit. For those datasets that are large and complex and not easily imported and fit into Phenobase, that information will be provided in a separate linked document rather than included in the data file itself.

Phenotypes: The parameters reported in datasets need to have names that describe their contents and their units need to be defined so that others understand what is being reported. A good name/label is short, unique (at least within a given dataset), and descriptive of the Phenotype. Column headings should be constructed for easy importing by various data systems. Choose a consistent format for each parameter and use that format throughout the dataset. All cells within each column should contain only one type of information (e.g. text, numbers, etc.). Common data types include text (alphanumeric strings of text), numeric, date/time, Boolean (also called Yes/No or True/False), and comments (for storing large quantities of text).

Coded Fields: Where the data value is not a discrete value (e.g. text or integers), they will be coded fields. Usually these will have standardised values but a key will always be provided to decipher the meaning of the code.

Missing Values: Leaving the value of a field blank is not an option in our data management processes as this poses a problem where a user might wonder if the data provider accidentally skipped a column. We use different codes to indicate different reasons why the data is missing (refer to Data Cleaning and Collection SOPs of the DTR).

Erroneous/Out of range Values: Similar to the missing values, a number of codes have been developed (format: 9999x) to indicate that there were some data reported/recorded but there was an issue with them (refer to Data Cleaning and Collection SOPs of the DTR).

Metadata

Metadata, defined as data about data, provides information on the identification, quality, spatial context, data attributes, and distribution of datasets, using a common terminology and set of definitions that prevent loss of the original meaning and value of the resource. Without descriptive metadata, discovering that a resource exists, what data was collected and how it was measured and recorded, and how to access it would be extremely difficult and prone to error. Metadata is particularly important to the DTR datasets as similar data is collected longitudinally over long periods of time from different projects. Often the equipment collecting these data or the method of asking the question changes slightly. Over time, without metadata, it would be impossible to harmonise these datasets to report them as similar data.

Data Standards

Data standards are required in order to describe objects, features, or items that are collected, automated, or affected by activities or the functions of organisations. At the DTR, data is carefully managed and organised according to defined rules and protocols. As we are a unit with a vast amount of data that is shared with and used by many research organisations all over the world, having clearly defined and followed data standards are particularly important (refer to Data Cleaning and Collection SOPs of the DTR). It is imperative that these data standards are continually monitored and updated in order to maintain compliance.

Benefits of adhering to our data standards include:

- More efficient data management (including updates and security)
- Easy data sharing
- Higher quality data
- Improved data consistency
- Increased data integration
- Better understanding of data
- Easier harmonisation across similar datasets and
- Improved documentation of information resources

Data Life-cycle Control

Managing the whole life cycle of data is an essential part of good data management. It includes:

- Data specification and structure, database maintenance and security
- Ongoing data audit,
- Archiving and backups to ensure data is maintained effectively, including periodic snapshots to allow rolling back to previous versions in the event of data corruption

Data Specification

As with computer programming or any complex design process, the majority of the work involved in building databases occurs long before using any software. Successful database planning takes the form of a thorough user requirements analysis, followed by data structure modelling.

The Phenobase has been designed by understanding user requirements ranging from data acquisition through data entry, reporting and long-term analysis. Data modelling is the methodology that identifies the path to meet user requirements. Our focus has been to keep the overall model and data structure as simple as possible whilst still adequately addressing the DTR's business rules and objectives.

By developing protocols and reviewing reference materials on our data we have defined entities, relationships and flow of information. This has always been an iterative and interactive process. The following broad questions have been answered:

- What are the Phenobase objectives?
- How will the Phenobase assist in meeting those objectives?
- Who are the stakeholders in the Phenobase?
- Who will use the Phenobase and what tasks do those individuals need it to accomplish?
- What information will the Phenobase hold?
- What are the smallest bits of information the Phenobase will hold and what are their characteristics?

- Will the Phenobase need to interact with other databases and applications? How will this be achieved?

Extra care has been taken in the database design and documentation in order to maintain data integrity during the lifetime of the Phenobase.

Database Maintenance

Technological obsolescence is a significant cause of information loss, and data can quickly become inaccessible to users if stored in out-of-date software formats or on outmoded media. Effective maintenance of digital files depends on proper management of a continuously changing infrastructure of hardware, software, file formats, and storage media. We expect major changes in hardware and software every 2-3 years. To deal with issues such as data security, minor changes can occur at any time and should be monitored on a regular basis. Datasets will continuously be migrated to new platforms and plans to make this as simple as possible are put in place. In a highly collaborative environment such as the DTR (Open Access Policy), versioning is extremely important and data users need to be made aware of what version of the data they are receiving at any point in time. Management of database systems requires good day-to-day system administration. Our database system administration is informed by a threat analysis so that it can employ means of threat mitigation and disaster recovery such as regular backups.

Data Audit

Good data management requires ongoing data audit to monitor the use and continued effectiveness of existing data and the DTR data team continuously monitors this. Our data or information audit involves:

- Identifying the information needs of the DTR and assigning a level of strategic importance to those needs
- Identifying the resources and services currently provided to meet those needs
- Mapping information flows between the DTR and internal analysts as well as external collaborators
- Identifying where changes are necessary by analysing gaps, duplications, inefficiencies, and areas of over-provision

An information audit not only counts resources but also examines how they are used, by whom, and for what purpose. The information audit examines the activities and tasks that occur in the DTR and identifies the information resources that support them and how they are used.

Benefits of a data audit include:

- Awareness of what data we hold assists with capacity planning
- Facilitate data sharing and reuse
- Monitor data holdings and avoid data leaks and duplication
- Recognition of data management practices
- Promote efficient use of resources and improved workflows
- Increase ability to manage risks – data loss, inaccessibility, compliance
- Enable the development/refinement of a data strategy

Data Storage and Archiving

Data storage and archiving address those aspects of data management related to the housing of data. This element includes considerations for digital/electronic data and information as well as relevant hardcopy data and information. Without careful planning for storage and archiving, many problems arise that result in the data becoming out of date and possibly unusable as a result of not being property managed and stored.

The DTR data to be stored includes:

- Files used by staff in their day to day office activities (typically small)
- Files used by analysts for their day to day activities (typically large)
- Files that will be shared with collaborators (typically via FTP)
- 2D and 3D photos of the cohort participants (numerous large files)
- Digitised documents such as questionnaires and survey responses (numerous)
- Personal and demographic data on the cohort participants (stored in a DBMS)
- Phenotypic data (stored in a DBMS)
- Phenotypic data (stored in a large repository as files)
- Genotypic and Expression data (stored in a large repository as files)

Physical dataset storage and archiving considerations for electronic/digital data include:

- *Server Hardware and Software* – What type of database will be needed for the data? Will any physical system infrastructure need to be set up or is the infrastructure already in place? Will a major DBMS product be necessary? Will this system be utilized for other projects and data? Who will oversee the administration of this system?
- *Network Infrastructure* – Does the database need to be connected to a network or to the Internet? How much bandwidth is required to serve the target audience? What hours of the day does it need to be accessible?
- *Size and Format of Datasets* – A rough maximum size of a dataset should be estimated so that storage space can properly be accounted for. The types and formats should be identified so that no surprises in the form of database capabilities and compatibility will arise.
- *Database Maintenance and Updating* – A database or dataset should have carefully defined procedures for updating. Due to the nature of a project its dataset may be live or ongoing and therefore will include such things as additions, modifications, and deletions. Versioning is important in this scenario.
- *Database Backup and Recovery Requirements* – A clear strategy for backing up any data/dataset should be in place in case of user error, software/media failure or disaster. Mechanisms, schedules, frequency and types of backups and appropriate recovery plans should be specified and planned. This can include types of storage media for onsite backups and whether off-site backing up (tapes) is necessary. From time to time the restoration process should be tested so that when required in a real situation, there are no nasty surprises.

Archiving of data is a priority in data management. Snapshots (versions) of data are created so that rollback is possible in the event of corruption of the primary copy and backups of that copy. As staff leave, the need to keep the data they created personally is ascertained through an “exit” interview with the data management team. Individuals storing data on their personal computers are required to make copies of their data on a more permanent and backed up, central storage unit. Projects that create data and are on one-off funding for the storage of data will archive their data on public repositories such as the EBI.

Longevity and Use

Data Security

Security involves the system, processes, and procedures that protect a database or file system from unintended activity. Unintended activity can include misuse, malicious attacks, inadvertent mistakes and access made by individuals or processes, either authorized or unauthorized. For example, a common threat for any web-enabled system is automated software designed to exploit system resources for other purposes via vulnerabilities in operating systems, server services or application.

Physical equipment theft or sabotage is another consideration. Accidents and disasters (such as fires, hurricanes, earthquakes or even spilled liquids) are another category of threat to data security. The data team stay current on new threats so that a database and file system and their data are not put at risk. Appropriate measures and safeguards are put in place for any feasible threats.
Classic methodology asks that security should be implemented in layers and should never rely on a single method e.g. uninterruptible power supply, mirrored servers (redundancy), backups, backup integrity testing, physical access controls, network administrative access controls, firewalls, sensitive data encryption, up-to-date-software security patches, incident response capabilities, and full recovery plans.

Risk Management

Risk is a function of the likelihood of a given threat-source‘s exercising a particular potential vulnerability and the resulting impact of that adverse event on an organisation. Risk management allows the data team to balance the operational and economic costs of protective measures with gains in mission capability by protecting the IT systems and data that support the DTR’s missions.
Risk management encompasses three processes:

- Risk assessment, risk mitigation and evaluation and assessment.

We use risk assessment to determine the extent of the potential threat and the risk associated with an IT system throughout its system development life cycle. The output of this process helps to identify appropriate controls for reducing or eliminating risk during the risk mitigation process.

Risk mitigation involves prioritising, evaluating and implementing the appropriate risk-reducing controls recommended from the risk assessment process. Because the elimination of all risk is usually impractical or close to impossible, it is the responsibility of the DTR senior management and the data team to use the least-cost approach and implement the most appropriate controls to decrease mission risk to an acceptable level.

The risk management process is ongoing and evolving. The information system will continually be expanded and updated, its components changed and its software applications replaced or updated with newer versions. In addition, personnel changes will occur and security policies are likely to change over time. These changes mean that new risks will surface and risks previously mitigated may again become a concern.

All DTR Servers are in a dedicated data centre behind lock and key with access only allowed for data and IT team staff. The databases and file servers sit behind the King’s College London firewall with strict rules as to access from outside the college network. Furthermore industry standard SQL server

DBMS are used to protect the Phenobase from unauthorised access. Transaction logs ensure that data can be rolled back to earlier points in time should any data corruption occur. Automated nightly backups of all databases are taken and stored on different file servers. All files are then backed up on a daily basis onto tapes. Monthly versions of these backups are physically stored in another location of the St. Thomas' Campus.

Data Access, Sharing, and Dissemination

Most of the DTR Data and information are available under an Open Access policy and are available to collaborators having requested the data and been approved by the Data Access Committee.

The following issues are addressed with regards to data access and our database system:

- Relevant data policy and data ownership issues regarding access and use of the data
- The needs of those who will require access to the data
- The costs of actually providing data as well as the cost of providing access to data
- Format appropriate for the data requestor
- Data requires restricted access to a subset of users
- Issues of private and public domain in the context of the data being collected
- Liability issues in terms of accuracy, recommended use, use restrictions, etc. We provide a disclaimer statement with each data provision so as to protect the DTR, the data collector or anyone associated with the dataset of any legal responsibility for misuse or inaccuracies in the data.
- Ensure anonymity of the participant where personal information is never provided to anyone. Only research data linked through an anonymous ID
- Sharing of certain sensitive data (death or cancer records) only with authorised and approved users

Whether certain data is made available or not and to whom is the decision of the TwinsUK executive Committee (TREC) as the data owner. Decisions to withhold or release data is based on the Open Access policy defined for the Wellcome Trust. Through weekly Data Access Committee meetings, the decision to withhold is totally transparent following the criteria in the stated policy.

All data shared with external collaborators is anonymised and access to personal information is given only to select members of staff have an absolute need to see/edit these (e.g. the Research Administration staff who recruit participants for studies or update their details).

Conclusion

Data management is essential for the effective usage of data. It has become even more important in the research field due to the vast amount of data available using electronic access. Using best practice methodologies the Department of Twin Research has:

- Defined policies, roles and responsibilities for data management
- Organised, documented, verified, and validated data to enhance its quality
- Managed the entire data life-cycle from design of a database to storage and archiving of data to disseminating data by providing appropriate access while maintaining security of the data.