H3ABioNet

Pan African Bioinformatics Network for H3Africa

Developing a Data Management Plan

H3Africa Data Management Workshop – 12th May 2016, Senegal

Introduction

- National Science Foundation definition of data: "data are any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data."

- Most funded biomedical research harnesses technological platforms to generate novel data for scientific analyses

- As "omics" technologies become more sophisticated, the volume of data generated has increased and costs have decreased – population based and large cohort studies at the individual genome level are now feasible

- These technologies generate large volumes of data and a significant amount of meta-data to enable context driven analyses (only if the data is well organized!!!)

Data Management

- Data management – the planned collection, quality control, storage, retrieval, analysis and dissemination of results helps to control information generated during a research project
- Managing data helps to control information generated in a structured manner
- Data management is an increasingly integral part of biomedical research, especially as datasets are getting much larger
- Biomedical sciences following a similar trend as physics where there are large datasets generated and teams of analysts around the world work on them

Data Management

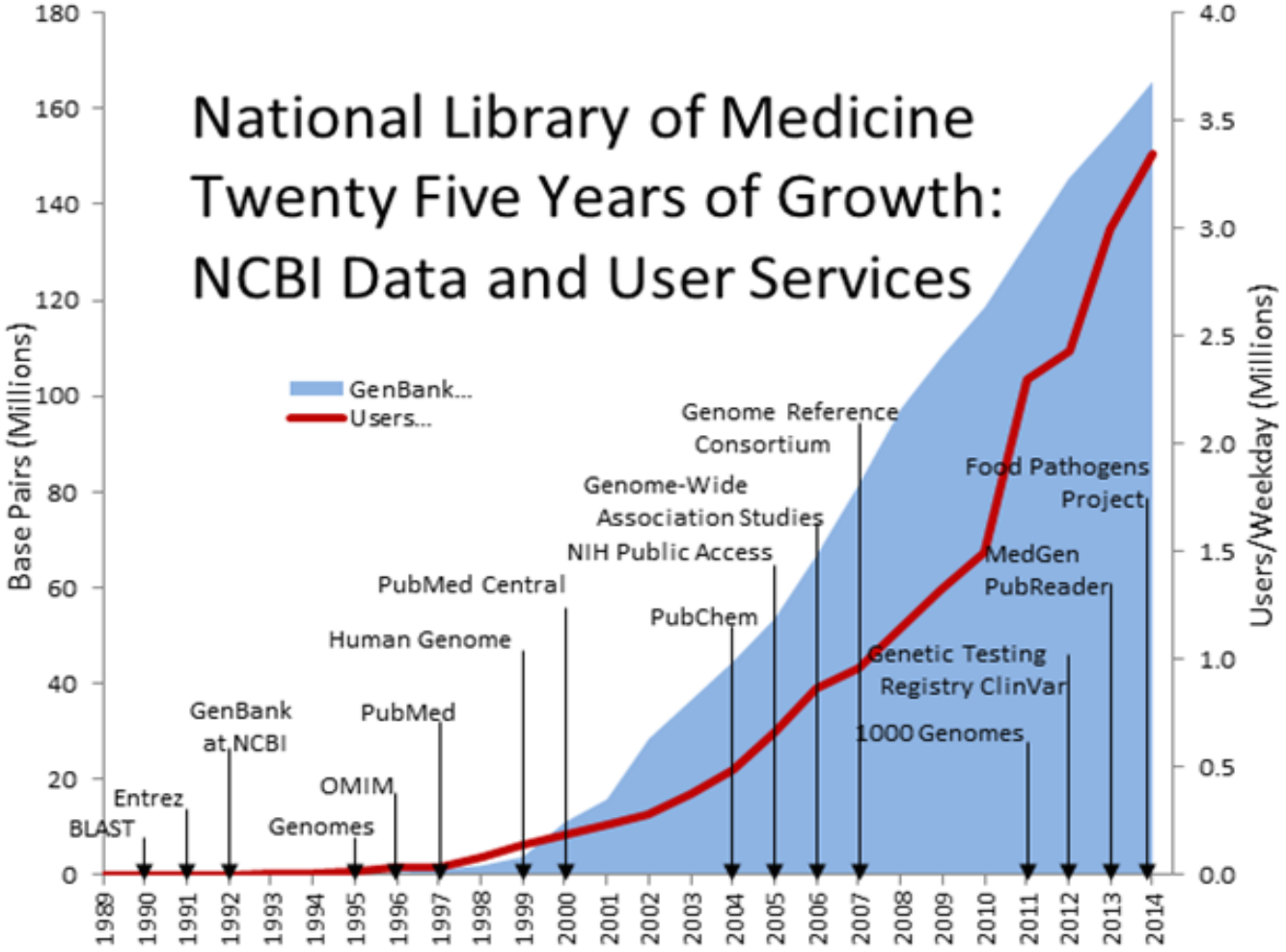


Figure from: https://www.nlm.nih.gov/about/2016CJ.html

Data Management Plan (DMP)

- A DMP is a document that describes how various aspects of research data generated for a project are handled during the project's life cycle and (especially in genomics) after the project is over

- The common elements in a DMP usually include*:
    1. Description of the data to be collected / created
    2. Standards / methodologies for data collection and management
    3. Ethics and Intellectual Property concerns or restrictions
    4. Plans for data sharing and access
    5. Strategy for long-term preservation

* http://www.dcc.ac.uk/resources/data-management-plans/faq-data-management-plans

Why care about a DMP?

- Funding agencies are recognizing the importance of research data as important community resources that accelerate the pace of discovery

- In some cases the data generated is funded by tax payers e.g NIH funded projects and hence should be readily available

- Other funding agencies such as Wellcome Trust recognize the Fort Lauderdale and Toronto statement on datasets that form community resources:
  - large-scale (requiring significant resources over time)
  - broad utility
  - creating reference datasets
  - associated with community buy-in

http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/index.htm

ADD THE NIH GRANTS DATA SHARING POLICY URL HERE!!!!!

Why care about a DMP?

- Most genomic (if not all) projects generate large datasets and when applying for grants – genomics projects will always most certainly need a DMP

- These DMPs do undergo peer review to evaluate the merit of a submitted grant proposal

- Although different agencies place varying levels on emphasis on the DMP, they will all definitely scrutinize your DMP and can revisit the DMP during performance reviews

- A DMP is an integral part for the success of a research project as the outcome is dependent on how well the data is managed

Why care about a DMP?

- Good data management can be challenging – especially when studies involve multiple sites, multiple PIs, site specific protocols, naming conventions, measurement scales and ethics

- The NASA Mars Climate Orbiter costing USD 125 million was lost in 1999 because two sets engineers working on different systems failed to convert metric units to imperial units

- A good DMP helps project affiliated researchers to:
  a. Define the various data attributes (metric or imperial)
  b. Easily find files for analyses
  c. Share and store their analyses with their multi-site partners
  d. Support the results of the published work

Ten Simple Rules for Creating a DMP

- **Rule 1**: Determine the Research Sponsor Requirements
- **Rule 2**: Identify the Data to Be Collected
- **Rule 3**: Define How the Data Will Be Organized
- **Rule 4**: Explain How the Data Will Be Documented
- **Rule 5**: Describe How Data Quality Will Be Assured
- **Rule 6**: Present a Sound Data Storage and Preservation Strategy
- **Rule 7**: Define the Project's Data Policies
- **Rule 8**: Describe How the Data Will Be Disseminated
- **Rule 9**: Assign Roles and Responsibilities
- **Rule 10**: Prepare a Realistic Budget

Rule 1: Determine funder's requirements

- Different funding agencies have different policies – some ask for specific details, others ask for broad plans
- Funding agencies usually provide DMP requirements in either the public request for proposals (RFP) or in an online grant proposal guide
- Keep in mind that the principle objective should be to create a DMP that will be useful for your project and should be treated as a living document
- Although funding agencies constrain the length of a DMP to a certain number of pages, a more detailed DMP can be submitted as an appendix or supplementary file

Rule 1: Determine funder's requirements

- Do find out the current DMP with regards to the funding agency by checking on their website for the latest version of documents e.g http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

- DMP templates for some funding agencies based in the U.K and the USA can be found at the DMPTool (https://dmptool.org/) and DMPonline (https://dmponline.dcc.ac.uk/) websites

- These tools are nice as they provide annotated advice for filling in the templates

- Do not be afraid to contact the responsible program officer for the RFP for more information and leave yourself ample time to write up a DMP!!

Rule 2: Identify research data to be collected

- **Research Data** - defines the data and materials generated and covered in a DMP

- **Types of data** – data and materials may take the form of physical samples (biospecimens), digital formats e.g genome sequences, electrocardiograms, clinical measurements (phenotypes)

- The types of data and their connections to each other should be defined and mapped before starting a DMP

- **Sources of data** – where is the data coming from e.g human participants, public databases or is it propriety data which can not be shared due to licensing

Rule 2: Identify research data to be collected

- Sources of data impact a DMP – if data is to be generated from human participants, ethics, privacy, security (sometimes IRB approval) needs to be included within the DMP

- Sources of the data will also inform sections within the DMP pertaining to intellectual property rights and ownership e.g the funding agency, the University

- **Volume of data** – the amount of data expected to be collected and stored, could be physical collections such as biospecimens or digital data

- Estimating the volume of data to be collected and stored is vital within your DMP as this will help define the budget for infrastructure and personnel costs

Rule 2: Identify research data to be collected

- If collecting 1 petabyte of data, ideal industry practice advices on having an onsite backup and an off site backup for disaster recovery

- That is in effect 3 petabytes of storage that needs to be budgeted for including personnel – in some cases impossibly slow to backup data so undertake replication which incurs internet costs

- **Data and file formats** - what formats will the data will be stored in as formats and versions change based on technology and archived data can become inaccessible due to legacy systems

- Determine common file formats widely used by your research community which are not heavily software / system or instrument reliant and have pre-defined standards e.g FatstQ, VCF, BAM or .csv, tab-delimited text using standard character encoding

Rule 3: Define how the data will be organized

- The types, sources and volume of data influences data organization e.g human participant information should be stored de-identified in a secure database system like REDCap or OpenClinica, and not as text files on a communal laboratory PC

- Software and strategies assisting with the organization, storage and analyses of data should be decided upon beforehand and included within the DMP

- Provision for human resources with skills to use these tools and undertake the organization of the data should be included in the DMP e.g if a graduate student is responsible then the data is lost / inaccessible once the student graduates and moves on

- If the data will be submitted to an access controlled repository it will need to be formatted and suitably organized for submission e.g EGA

Rule 4: Explain how the data will be documented

- **Metadata** – additional data that provides contextual details of research data generated

- Provides standardized, structured information on how the actual data was collected, processed and interpreted e.g measurement types, instruments used, analytical methods, software versions

- Good data documentation adds value to datasets and the reuse of the data for other studies is directly related to the amount and quality of metadata associated with it

- Important in the context of meta-analyses where multiple datasets can be merged to conduct analyses previously impossible due to small sample sizes or being underpowered

Rule 4: Explain how the data will be documented

- Good data documentation should include the following*:
  - Data listing with descriptions for cases, individuals or items studied or collected
  - Names, labels and descriptions for variables, records and their values
  - Explanation of codes and classification schemes used
  - Codes for missing values and reasons for values missing
  - Derived data created after collection, with analyses packages / software and the versions used provided

- In a lot of cases the metadata standards and schemas have been defined for various communities e.g ontologies

- **Ontology** – common vocabulary used to describe various data attributes by a community that needs to share data

Rule 4: Explain how the data will be documented

- Ontologies differ from controlled vocabularies / data dictionaries as the definition of terms and their hierarchical relationships are explicitly defined e.g “is part of”

- Controlled vocabularies have explicit definitions of terms but in most cases relationships between terms have not been mapped

- Significant development of ontologies in the biomedical community e.g Gene ontology (GO), Experimental Factor Ontology (EFO), Human Phenotype ontology (HPO)

- Determine if an ontology is available that can be used to describe the data: https://bioportal.bioontology.org/projects

Rule 4: Explain how the data will be documented

- More then one ontology might be needed in some cases e.g EFO to describe sequencing instrument data and HPO to describe the phenotypic features of human hereditary and other diseases

- Where there are no ontologies available, widely used community adopted controlled vocabularies should be used

- Defining the types of data assists in determining what metadata schemas are available and appropriate for use

- Use of ontologies and community adopted controlled vocabularies assists in the searching and finding of data and are required by most data repositories

Rule 5: Describe how data quality will be assured

- Some RFPs do indicate what requirements for Quality Controls / Quality Assurances are expected

- Depending on the nature of the study and the types of data being collected and /or generated, the focus on each of the above activities within the DMP will differ

- QC / QA of data within a DMP could encompass:
  - Data collection procedures that minimize variation
  - Lab instrumentation used (or certifications and summary of QC / QA processes used if an external vendor is used for data generation)
  - Transportation, processing and storage of samples
  - Data capture procedures
  - Data analyses procedures
  - Continuous training of staff to ensure good laboratory practice or certified training of staff to use new equipment

Rule 5: Describe how data quality will be assured

- The QC / QA of data fidelity is linked to the data types being collected and also for analysis
- Interms of data entry, verification and the use of specifically constrained fields should be used e.g “@” for a valid email address, range of dates
- Other QC could be physical sample inspection, testing for purity
- In the case of analyses each type of data analyses does have its own QC steps e.g trim bases that have Phred scores less then 35
- Visual inspection of the data during analyses helps in the QC process e.g QQ plots in GWAS to determine confounding effects within population structure between cases and controls

Rule 6: Present a sound data storage and preservation strategy

- Collecting, generating and analysing data incurs a significant amount of cost and personnel time

- A lack of data storage and preservations strategy could see 3 – 4 years of work lost in the amount of time it takes for a computer to crash or a hard drive failure

- A DMP should take into account:
  - How long should the data be stored and available for
  - How will the data be secured and shared amongst collaborators before being made available to the wider community
  - How the data will be archived and made retrievable in future

Rule 6: Present a sound data storage and preservation strategy

- Data will need to be stored while being collected and analyzed which should be estimated beforehand from the actual grant application
- Provision for the storage of data in two locations and one off site location for disaster recovery should be articulated
- Measures to protect the data should be included – the level of data security and access is dependent on the nature of data
- These security measures should include provisions on how off site project members / collaborators access the data e.g multiple copies of the data in multiple locations increases the chances of a data security breach

Rule 6: Present a sound data storage and preservation strategy

- Once the analyses is completed and results published, the data might need to be made publically available according to the funding organization's policies

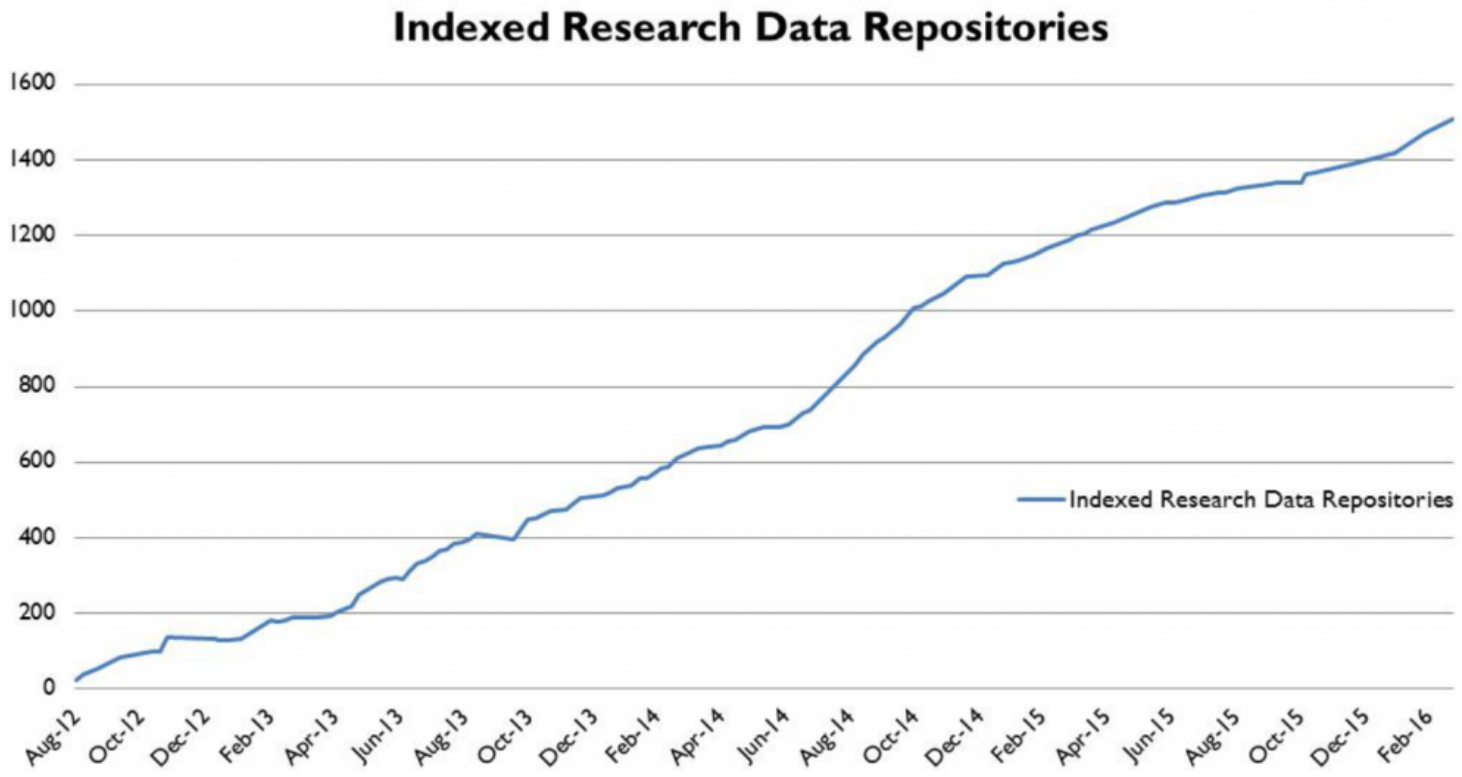
- Archiving, long term storage and retrieval of data upon request or application is an expensive endeavor that incurs significant costs in infrastructure and personnel to manage the infrastructure over a sustained period

- The data storage solution provided should include plans for scalability incase more data then is anticipated ends up being collected

- Usually University IT departments are not well equipped to do this as they follow a cost recovery model

Rule 6: Present a sound data storage and preservation strategy

- Funding allocated for storage and personnel costs to maintain the infrastructure ceases once a grant ends
- Hence data repositories are increasing in terms of their diversity and volume of data being stored (http://www.re3data.org/)



Rule 7: Define the project's data policies

- Many funding organizations require that DMPs include explicit policy statements about how data will be managed and shared

- Such policies should include if appropriate:
  1. Licensing or sharing arrangements that pertain to the use of preexisting materials
     - Ideally demonstrate that advice has been sought on anticipated resources and address any copyright and / or licensing issues
     - Identify and include a description of the relevant licensing and sharing arrangements in your DMP
     - Usually useful to approach a University research officer and relevant Institutional bodies

Rule 7: Define the project's data policies

2. Plans for retaining, licensing, sharing, and embargoing (i.e., limiting use by others for a period of time) data and other materials
    - Data ownership should be clarified and be inline with Institutional policies and funding organization
    - Explain how and when the data and other research products will be made available
    - Indicate if there might be any potential restrictions or barriers to data sharing such as the need to safe guard the research participants, applying for intellectual property protection

3. Legal and ethical restrictions on access and use of human subject and other sensitive data
    - Be explicit about the type of consent obtained, de-identification or annonymisation of participants
    - Depending on the funding organizations, Institutional Review Board ethics approval might be needed

Rule 8: Describe how the data will be disseminated

- Providing access to research data enables the reuse of data and adds value
- A plan for the disseminations of data is usually required by the funding organization
- The data dissemination plan should be as concise and specific as possible stating how, when and what data will be released
- A data dissemination plan that has the minimum possible restrictions to the release and access of data when the project is completed and results published is the most preferred where applicable

Rule 8: Describe how the data will be disseminated

- Funding organizations do provide a list of public repositories based on the data to be submitted e.g DNA, protein structures
- Certain types of data of the same study can be submitted to different repositories
- E.g sequence data from a microbiome study can be submitted to the publically accessible European Nucleotide Archive (ENA) while the phenotype data is submitted to European Genome Archive (EGA)

Rule 9: Assign roles and responsibilities

- A comprehensive DMP should clearly designates the roles and responsibilities of every named individual and organization associated with the project

- Roles may include data collection, data entry, QA / QC, metadata creation and management, backup, data preparation and submission to an archive, and systems administration.

- Large multi-investigator projects may benefit from having a dedicated staff person(s) assigned to data management

- Treat your DMP as a living document and revisit it frequently (e.g., quarterly basis). Assign a project team member to revise the plan, reflecting any new changes in protocols and policies.

Rule 10: Prepare a Realistic Budget

- A common error when developing a DMP is forgetting to budget for all activities involved in the data management life cycle for the activities
- Data management is time consuming, costs money in terms of software, hardware, and technically skilled personnel that are highly sought after by industry
- Review your DMP and make sure to link components in your DMP with specific line items with the budget proposal and the budget justification to support the people that manage the data as well as pay for the requisite hardware, software, and services
- Check with your IT department and the preferred data repository department so that requisite fees and services are budgeted appropriately

The DMPTool

- The DMPTool is a collaboration of multiple institutions, including DataONE, and is a service of the UC Curation Center. The DMPTool will help you:
- Create ready-to-use data management plans for specific funding agencies;
- Meet funder requirements for data management plans;
- Get step-by-step instructions and guidance for your data management plan as you build it;
- Learn about resources and services available at your institution to help fulfill the data management requirements of your grant.

Conclusions

- A good DMP takes a lot of thought and engagement with various stakeholders such as the University's IP office, IT department, financial officer – start early

- At first, developing a DMP is hard work and time consuming, but once done it can be reused with information updated from various sources that will enable the subsequent DMPs to be written up relatively quickly with most successful parts of the DMP being incorporated into subsequent projects

- A data management plan should provide your project members, funders and others with an easy-to-follow road map that will guide and explain how data are treated throughout the life of the project and after the project is completed

Conclusions

- A DMP provides a vehicle for conveying information to and setting expectations for your project team during both the proposal and project planning stages, as well as during project team meetings later, when the project is underway.

- The best plans are “living documents” that are periodically reviewed and revised as necessary according to needs and any changes in protocols, policy, technology, and staff, as well as reused, in that the

Break out session instructions

1. Organize yourself into groups that are part of the same H3Africa project

2. Appoint a rapporteur for your group to report back to the workshop group

3. Pick the project you are affiliated to from: http://h3africa.org/consortium/projects

4. Click on your affiliated project weblink to obtain the project's summary

5. Go over the data management template, discuss the various sections amongst yourselves and complete the various sections of the DMP template using your affiliated H3Africa project as a case example

6. Appoint a rapporteur from your group that will provide feedback on the DMP section discussions from the DMP template

Useful URLs

http://www.nature.com/nature/journal/v527/n7576_supp/full/527S16a.html

http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/index.htm

http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf

http://www.dcc.ac.uk/resources/data-management-plans/faq-data-management-planshttp://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How%20to%20Develop.pdf

http://www.lshtm.ac.uk/research/researchdataman/plan/wellcometrust_dmp.pdf

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP-checklist-flyer.pdf

https://www.libraries.psu.edu/psul/pubcur/what_is_dm.html#what-is-data-management