Journal of Biomedical Informatics 122 (2021) 103900

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Special communication

Data Management Plans in the genomics research revolution of Africa: Challenges and recommendations

Faisal M. Fadlelmola [a,*], Lyndon Zass [b,1], Melek Chaouch [c,1], Chaimae Samtal [d,1], Verena Ras [b,1], Judit Kumuthini [e], Sumir Panji [b], Nicola Mulder [b]

[a] Centre for Bioinformatics and Systems Biology, Faculty of Science, University of Khartoum, Al-Gamaa Ave, Khartoum 11115, Sudan
[b] Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, CIDRI Africa Wellcome Trust Centre, University of Cape Town, South Africa
[c] Laboratory of Bioinformatics Biomathematics and Biostatistics (LR16IPT09), Institut Pasteur de Tunis, 13 Place Pasteur, B.P. 74 1002 Tunis, Belvédère, Tunisia
[d] Laboratory of Biotechnology, Environment, Agri-food and Health, Faculty of Sciences Dhar El Mahraz–Sidi Mohammed Ben Abdellah University, Fez 30000, Morocco
[e] South African Bioinformatics Institute (SANBI), University of Western Cape (UWC), Life Sciences Building, Bellville, Cape Town, South Africa

ARTICLE INFO

Keyword
Data Management Plan
Africa
FAIR
Funders
Data sharing
Genomics data management

ABSTRACT

Drafting and writing a data management plan (DMP) is increasingly seen as a key part of the academic research process. A DMP is a document that describes how a researcher will collect, document, describe, share, and preserve the data that will be generated as part of a research project. The DMP illustrates the importance of utilizing best practices through all stages of working with data while ensuring accessibility, quality, and longevity of the data. The benefits of writing a DMP include compliance with funder and institutional mandates; making research more transparent (for reproduction and validation purposes); and FAIR (findable, accessible, interoperable, reusable); protecting data subjects and compliance with the General Data Protection Regulation (GDPR) and/or local data protection policies. In this review, we highlight the importance of a DMP in modern biomedical research, explaining both the rationale and current best practices associated with DMPs. In addition, we outline various funders' requirements concerning DMPs and discuss open-source tools that facilitate the development and implementation of a DMP. Finally, we discuss DMPs in the context of African research, and the considerations that need to be made in this regard.

1. Introduction

The African continent faces a large burden of both communicable and non-communicable diseases [1]. Individuals of African descent are known to possess high genetic diversity, which may influence susceptibility and resistance to, as well as treatment of these diseases [2,3]. To investigate the link between genetics and disease, African researchers are increasingly performing genome-sequencing studies that generate large datasets. The Human Heredity and Health in Africa (H3Africa) consortium, funded by the National Institutes of Health (NIH) and the Wellcome Trust (WT), through AESA (Alliance for Accelerating Excellence in Science in Africa), support several such studies and have thus far generated genome sequence and/or variation data for over 50,000 African individuals [4]. The H3Africa bioinformatics network (H3ABioNet) aims to develop bioinformatics capacity and provide infrastructure to support genomic analysis and data storage throughout the continent [5,6]. The increasing number of genomic studies in Africa, have also led to an increasing number of training opportunities, which facilitate genomics data analysis, computational skills, and data management.

Data management is an important part of any large-scale research process, particularly one that involves genomics- and related platforms. To keep track of data generated in such studies, as well as effectively organize and maintain it, a Data Management Plan (DMP) is essential. A DMP is a document that is developed before the start of a research project and details how research data will be collected, generated, and processed during the project, as well as stored and shared thereafter. DMPs have become a core component of modern academia due to the increasing amount of Big Data studies being conducted worldwide. The benefits of writing a DMP are multifold, including: 1) allowing compliance with funder mandates; 2) enabling transparent research; 3) enhancing data FAIRness (findability, accessibility, interoperability, and reusability); 4) protecting data subjects; and 5) allowing compliance

* Corresponding author.
E-mail address: faisalfadl@gmail.com (F.M. Fadlelmola).
[1] These authors contributed equally to this manuscript.

https://doi.org/10.1016/j.jbi.2021.103900
Received 12 March 2021; Received in revised form 24 August 2021; Accepted 25 August 2021
Available online 8 September 2021

with local data protection policies [7,8]. As research collaboration has become the norm in modern research, data sharing, and secondary data access are increasingly relevant topics and important to cover in a DMP. Secondary access to research data increases the returns from public investment in the field, reinforces open scientific inquiry, encourages diversity of studies and opinion, promotes new areas of research, and enables the exploration of topics not envisioned by the initial investigators. For these reasons, the Organization for Economic Co-operation and Development (OECD) developed guidelines to facilitate cost-effective access to digital research data from public funding [9]. In addition, guidelines for developing a good quality DMP have also previously been described [7].

In this review, we highlight the importance of DMPs in modern biomedical research, explaining both its rationale and current best practices. In addition, we discuss various funders' requirements with respect to DMPs and provide brief descriptions of existing open-source instruments that encourage advancement and usage of a DMPs. We also assess and examine DMPs within the setting of African research environment, and touch on the best practices and measures that ought to be made in this respect.

2. Rationale for a DMP

One of the primary reasons to develop a DMP is due to the requirement outlined by funding agencies during the funding application phase. Two of the key outcomes within the research enterprise are publications and data products [7]. Data generated by federally funded projects are also seen as publicly funded data and, therefore, funders often require the sharing of such data for future use. A DMP can thus be seen as a blueprint for the management of data throughout the data lifecycle [10]. It provides evidence that the funding applicant has a proper grasp on the amount and significance of the data that will be produced during the project. Data that has been collected or produced following community standards, which have been stored appropriately, and with comprehensive metadata often lead to results that are more meaningful [7]. Historically, many DMPs stemmed directly from the need to support reproducibility [11]. A well-drafted DMP considers the standards; descriptors and metadata that will be implemented in a research project, as well as how the data will be stored and shared, and is thus crucial for the longevity of your data, facilitating:

1. Data arrangement, composition, and flow, flagging areas where assistance may be required ahead of data collection or generation, and thus limiting mistakes. This too maintains a strategic distance from pointless data duplication.
2. The archiving of the data lifecycle, allowing researchers to think about the conservation and sharing of the data. Planning data sharing and secondary data use can also facilitate the development of consent forms employed during a research project.
3. Reproducibility, because it reduces the probability of disasters such as data misfortune, mistakes and unscrupulous utilization of data. It also helps to track provenance and how the data are processed.
4. The production of high-quality data, with appropriate metadata.
5. The arrangement of data security and data assurance approaches required.

Good management of data could also lead to data publication, which can be seen as an alternate form of dissemination and is gaining popularity amongst researchers [11], while data management in and of itself is becoming an acknowledged research skill.

DMPs are developed to maintain FAIR, accurate and useful content; this should lead to data being efficient, well-managed in the present, and prepared for preservation in the future. As shown in Fig. 1, thinking ahead is vital, as considering and planning data management at the beginning of the project will help compliance with requirements of most research funders as an integral part of the grant application, and/or developed iteratively during the research activity. However, a good DMP will develop as your research progresses, reflecting the changes in your research and technology offering.

3. Current best practices for research data management policies

An increasing number of academic institutions are creating or adopting Research Data Management (RDM) Policies. Although these policies often share common themes, such as data ownership, retention and stakeholder responsibilities, RDM policies may also differ quite extensively between organizations [12]. The development of RDM policies is often led and supported by university libraries, as they most likely already possess metadata and archiving services [13]. With most universities opting to include libraries in their RDM policies, it is recommended that anyone endeavouring to develop their own DMP involve library representatives from their respective institutions [14]. Since a growing demand for DMPs exist, several resources have been developed to assist those wanting to develop their own plan. Many templates have been developed along with several tools to assist researchers with assessing their data management needs and capabilities, such as the

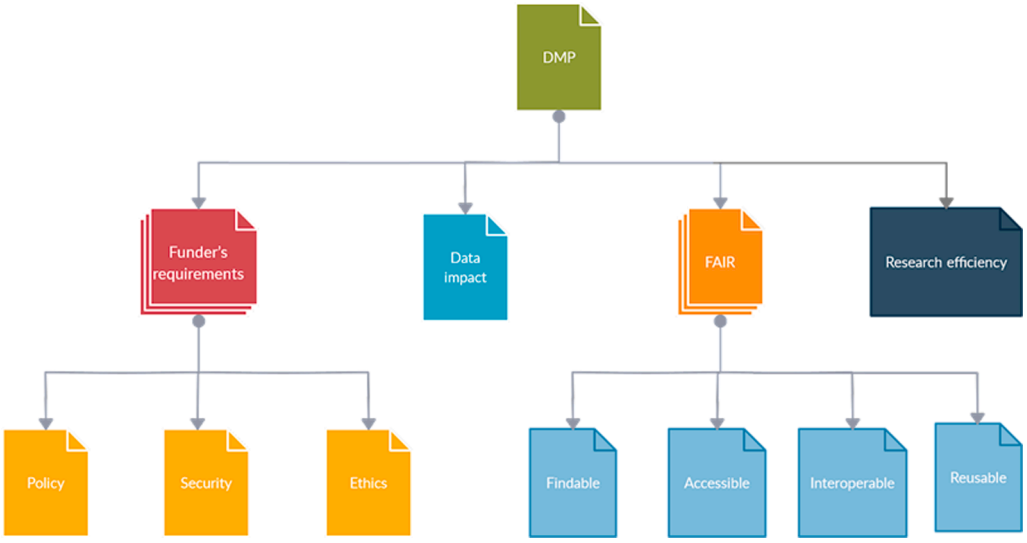

Fig. 1. Research data management advantages.

RDM readiness survey and the Research Data Alliance (RDA) standard for machine-actionable DMPs [15–19]. Machine-actionable DMPs are open, shared and interoperable concepts which facilitate data discovery and reuse, and enable automated evaluation and monitoring of large datasets [19,20]. The genomics research community in particular has benefitted from standardization of machine-actionable DMPs and common practices and strategies associated with data management to enable future work in this area [21].

A number of DMP best practices have been recommended by data one.org; and these are further discussed below, supplemented with recommendations from Parker et al., who provided the first major human genomic data archiving service in Africa [22].

During **project planning**, the project aims and objectives should be outlined and clarified in order to determine the types of data that will be collected and/or generated. This will allow researchers to identify community standards or methodologies relevant to the collection and generation of such data and provide a rough estimate of the data management needs (volume and breadth). Once identified, a data manager, whose responsibilities should be clearly outlined in the DMP, may be required to oversee the process. Some best practices to consider during **data collection and generation** include: 1) how data **quality control and validation** will be performed and how data that fails validation will be treated; 2) what **metadata** will be collected and how – does it adequately describe the data; 3) the use of **standard terminologies** (e.g. controlled vocabularies or ontologies); and 4) the development of a data dictionary which provides a detailed description of every element in the dataset, assisting downstream use and reuse of the data.

Deciding where and how data will be stored before a project commences is crucial to prevent data loss and corruption, however, a process to follow in the event of data loss should also be detailed in the DMP. Some best practices to consider during **data storage** include: 1) the **storage system** - it is best to create, document and manage your storage system, employing version control; 2) it **data backup strategies or policies**. Here you may outline who has access to the backup, where, how and when, for example; 3) determining what data will be preserved to accurately plan and assess storage needs; and 4) deciding early on how missing data will be treated.

Data responsibility, accountability and authority are crucial to consider when developing the **data sharing** portion of the DMP. These considerations are largely captured under data protection policies such as the General Data Protection Regulation (GDPR) policy in Europe. In an era of Big Data science, driving innovation whilst protecting the rights of the individuals behind such innovation is key, therefore data protection policies have been adopted globally to protect research participants and data consumers. Within the genomics field this is particularly relevant because individual genomes are considered personally identifiable information – such as name and date of birth. Personally identifiable information can be used to exploit individuals or cause unjust stigmatism and discrimination. Therefore, it is of utmost importance to share genomics data responsibly, ensuring that third parties use it with scientifically and socially justified intentions, particularly with regards to African research participants and other historically disenfranchised minorities. When considering the relevance of data protection policies to a specific research project, a DMP enables researchers to consider the aforementioned in advance, and cover topics such as: 1) **data sensitivity** – determine whether you have any sensitive data within your dataset (such as personally identifiable information) and how it will be treated. This is normally governed by the institutional policy and/or national legislation; 2) **data access** – determine with what level of access data will be shared, such as controlled access (employed by H3Africa); 3) **data ownership** – identify early on who owns the rights to the data. This may be a combination of parties such as the principal investigator, research institution and funder; 4) **repository** – identify suitable repositories for submission before commencing data collection, shaping a DMP toward a repository will increase the likelihood of acceptance to that repository; 5) **licensing and copyright** – this typically involves assigning a unique identifier to your dataset that will assist in data discovery and any legal considerations around the re-use of your data. This also facilitates data provenance. This can be a rather complex topic as it must adhere to the requirements of funders and the institution, but must consider the limitations of the consent signed by study participants.

The increasing requirements to share data has created a growing concern amongst funders and repositories that current funding models will prove inflexible and not meet the growing demands for sharing and storing data. Nevertheless, this demand has also resulted in some innovation amongst repositories, both commercial and non-commercial. Repositories such as Dryad [https://datadryad.org/stash] traditionally shared only datasets associated with a published paper, but they recently developed a platform that allows for sharing of standalone datasets. Likewise, the H3Africa consortium has developed a data archiving service to assist their many projects with submission of data to the European Genome-Phenome archive (EGA), while retaining a local copy of that data to ensure its retention and availability on the African continent [22,23].

4. Limitations of Data Management Plans

A potential limitation of DMPs which vary between funders is their attention to the management and sharing of data post-publication [24,25]. Specific details regarding the data collection, processing, analysis, formats, collaborative/consortium sharing policies and data stewardship practices that need to be put in place are not required within a DMP at the time of a grant application submission [24]. This is understandable as a comprehensive DMP with specific details before a project is even considered for funding is hard to implement and requires expertise developed over an extended period. Additionally, the use of community accepted repositories for data submission helps to ensure that some requirements for metadata standards, preservation and accessibility or sharing of data post-publication are met [24].

However, the utility of a DMP is less effective if not treated as a living document that is periodically reviewed and updated as a research project progresses. Some funding agencies do not require updated or reviewed DMPs to be submitted annually as part of a progress report. In academic research environments, which usually have a high personnel turnover as students graduate and post-Doctoral researchers move on, updating of DMPs to include information about the collection and handling of data is rarely done. To some extent, this can be done at the time of data submission by employing standards to capture experimental provenance and enable reproducibility such as the Minimum Information about a Microarray Experiment (MIAME) standard [26]. Keeping a DMP updated requires a time commitment from researchers which subtracts from time devoted to research and hence, tends to be lower on a list of priorities with little incentive. Investment in technology and personnel for good data stewardship competes with using as much funding as possible to generate data and undertake research to produce results, and the time and effort required for good data curation and stewardship is often underestimated [27].

DMPs as tools are only as useful as the people that use them. As new requirements for data stewardship and implementation of policies such as FAIR and GDPR have come into effect, training on how to incorporate these into DMPs need to be provided to new generations of researchers [24–26]. It is difficult for a single institution (usually located within Digital Library Services) to provide domain specific training to specialized researchers e.g. structural biologists or human population genomics researchers that would use different file formats, repositories, meta-data standards and software, even though they could be located within the same faculty, especially in resource constrained environments.

5. DMP requirements by funders

Several funders currently require the sharing of research data to the

wider research community, and therefore recommend the development of a DMP when submitting a research proposal. Some of these funders are further discussed below and summarized in Table 1 [28]. Many of the themes highlighted by these funders have also been supported by crosscutting, non-funding organizations such as the World Health Organization (WHO) and the African Open Science Platform (AOSP). Similarly, the OECD have made several recommendations for low- and middle-income countries supporting data management.

5.1. NSF

The National Science Foundation (NSF) is an independent agency of the US government that supports fundamental research in science and engineering. The NSF offers very general guidelines for DMPs, and each directorate and/or program offers guidelines that are more specific. A universal guideline was deemed inappropriate because each discipline may have varying criteria for research data and different expectations about data storage and sharing. The NSF empowers peer-reviewers and program managers to set the standards for data management within the relevant "community of interest."

5.2. NIH

The NIH is the medical counterpart to the NSF, responsible for biomedical and public health research. Several data management considerations are highlighted by the NIH. In terms of data sharing and preservation, data sharing and archiving costs are considered research costs, and researchers are expected to find a balance between the cost and value of data preservation. Moreover, data supporting the main findings of a research paper should be released no later than the date of publication acceptance to facilitate secondary use, and access to this data is expected to be free [29].

5.3. WT

The WT is a charity organization which urges researchers to carefully consider their DMPs during the project proposal phase. WT requires researchers to maximize the availability of any results produced by a study provided these results could have significant value as a resource for other researchers. They also explicitly state that as few restrictions as possible should be placed on the data and open access is preferred. The WT requires the data supporting the main findings of a publication as well as any original software developed to analyze that data, be made publicly available at the point of publication. They also require an "outputs management plan" be made available at the proposal stage. This plan must address the following key areas: data and software outputs; research materials; intellectual property and required resources (may also refer to skilled people). Since WT is a charity organization, should there be a health emergency, WT-funded researchers are required to immediately make their final results available, even when they have not yet been published.

Table 1
DMP Funders' requirements*.

● Full Coverage: All requirements applied
◖ Partial Coverage: All requirements not fully applied
○ No Coverage: Requirements not applied

	Datasets policy	Sharing/ access	Time limits	Curation	Monitoring	Guidance	Repository	Published outputs
NIH	●	●	●	●	●	●	●	●
WT	●	●	●	●	●	●	●	●
NSF	●	●	●	○	◖	◖	●	●
EC (H2020)	●	◖	◖	◖	◖	●	●	●

- **Datasets policy**: a datasets policy or statement on access to and maintenance of electronic resources
- **Access/sharing**: promotion of Open Access journals, deposit in repositories, data sharing or reuse
- **Time limits**: set timeframes for making content accessible or preserving research outputs
- **Curation**: maintenance and preservation of research outputs
- **Monitoring**: whether compliance is monitored or action taken such as withholding funds
- **Guidance**: provision of financial assistance
- **Published outputs**: a policy on published outputs e.g. journal articles and conference papers

* Adapted with major additions and modifications from the Digital Curation Centre [28].

5.4. MRC

The Medical Research Council (MRC) is a national funding agency established in various countries, dedicated to improving human health by supporting research across the entire spectrum of medical sciences. All research data generated through MRC-funded research must be managed and curated effectively throughout its lifecycle, including archiving, to ensure integrity, security and quality, and, where possible, to support new research and research data sharing to maximize the benefit and impact of MRC research funding [30,31]. Records should be kept to enable understanding of what was done, how and why, and to allow the work to be assessed retrospectively and reproduced if necessary. Information relating to participant consent should be held securely and subject to the same retention criteria as the primary/raw data [30].

5.5. BBSRC

The Biotechnology and Biological Sciences Research Council (BBSRC) is part of the UK Research and Innovation partnership. It sponsors a wide range of scientific research that generates large volumes of data. A concise plan for data management and sharing is required, which may include details of data areas, types and formats, standards and metadata, secondary use, methods for data sharing, and timeframes for release. Guidance is available under 'Data Sharing Areas' on p7 of the BBSRC Data Sharing Policy (https://www.bbsrc.ac.uk/documents/data-sharing-policy-pdf/).

5.6. Bill and Melinda Gates Foundation

The Bill & Melinda Gates Foundation is another charity organization committed to optimizing the use of health-related data to translate knowledge into life-saving interventions. Therefore, it is essential that grantees make data widely and rapidly available to the broader global health community through good data access practices, to promote innovation, collaboration, efficiency, accountability and capacity strengthening. The foundation is driven by a number of data access principles, and as such, requires the development of a data access plan which includes the nature and scope of data to be disseminated as well as the timing thereof, the manner of data storage and dissemination, and access conditions.

6. Data management tools

Numerous tools are currently available to facilitate both the development and implementation of a DMP. The majority of these tools focus on the management of the research data collected and produced during the course of a project. Some of these resources and their main features are listed in Table 2. When considering which tools to use for development and implementation of a DMP, it is useful to refer to the 10 principles for machine actionable DMPs previously described, that may enable parts of the DMP to be automatically generated and shared (e.g., with collaborators and funders) [32]. In addition to the tools listed, The Library Carpentries and Australian Research Data Commons published a list of "FAIR Data and Software Things" [33], outlining how various research components can be made FAIR and providing resources to facilitate FAIR compliance.

7. DMPs in the context of Africa/H3Africa projects

There are numerous challenges associated with implementing DMPs in practice, and some of these challenges have previously been raised by

Table 2
Data Management Tools for Planning and Development.

Tool Name	Main features	Reference
DMPTool (https://dmptool.org/)	– Facilitates DMP development for projects and proposals. – Continuous software improvement – Several DMPs made using the software available to the public (https://dmptool.org/public_plans).	NA
The Integrated Rule-Oriented Data System (iRODS) (https://irods.org/)	– Supports collaborative research, management, sharing, publication, and long-term preservation of data. – Allow scaling to collections containing petabytes of data and hundreds of millions of files.	[34]
REDCap (Research Electronic Data Capture) (https://irods.org/)	– Secure web application for building and managing online surveys and databases. – Online or offline project design, availability, secure and web-based, multi-site access, fully customizable, audit trails, automated export procedures. – Community continuously aims to develop the software in order to address evolving user needs.	[35]
DMPOnline (https://dmponline.dcc.ac.uk/)	– Web-based tool to help researchers create, review, and share DMPs that meet institutional and funder requirements. – Provides step-by-step guidance and other information. – Includes a number of templates for funders.	NA
ezDMP (https://ezdmp.org/)	– Web-based tool, free to all investigators. – Facilitates development of DMPs for NSF Grant applications. – Includes links to updates from the Directorate of Biology on DMPs as well as a list of biology-specific repositories.	NA
ARGOS (https://devel.opendmp.eu/home)	– An online open extensible service that simplifies management, validation, monitoring and maintenance of DMPs. – Allows creation of actionable DMPs that may be freely exchanged among infrastructures. – Provides a flexible environment and easy interface for users to navigate and use.	NA
DMPlanner (https://dmplanner.athenarc.gr/)	– Web application for creating DMPs – Extracts information from ORCiD and public repositories the users used for their project.	NA
RDM Toolkit (https://rdmtoolkit.jisc.ac.uk/)	– Curates a wide range of resources including websites and tools to support RDM. – Developed an online, interactive research data lifecycle, outlining the major data management steps. – Provides useful guides, training materials, tools and resources to aid DMP development. – Training materials are signposted and classified according to three main audiences (researcher, support staff and IT).	NA
Open Research Hub (https://www.jisc.ac.uk/open-research-hub)	– An interoperable system for managing, preserving and sharing institutional digital research data. – Three service options including, end-to-end, repository and preservation services. – Digital files automatically undergo preservation on deposit and allow mediated deposit of data. – Reporting tool that generates reports based on information within your own system and any other integrated services.	[36]
AnVIL (https://anvilproject.org/)	– Interoperable cloud-based resource. – Co-locating data, storage and computing infrastructure with commonly used services and tools for analyzing and sharing data. – Data access and data security.	[37]

Lefebvre et al., including funding concerns, lack of RDM knowledge, and a lack of specific guidelines on how to implement DMP concepts such as making data FAIR (25). Because African researchers face many unique challenges not experienced elsewhere (e.g., internet connectivity, infrastructure, capacity), there are several special considerations that need to be taken into account when drafting and implementing a DMP in Africa. The H3Africa consortium, as a driver of large-scale genomics research projects in Africa, has had to overcome many such challenges. Some of these challenges and how H3Africa have managed to overcome these challenges, are briefly described here.

7.1. Data collection, generation and standardization challenges

There is a need for data collection to be standardized across the research community, particularly so in biomedical research, however, existing data collection standards are not always appropriate for use in African research settings, therefore provision for such adjustments need to be made. An appropriate balance between using well-established existing standards and adapting them needs to be found. Similarly, there may also be a need to make certain adjustments for language differences. Africa is linguistically diverse, and certain English terminology is understood differently in the African context. Moreover, various regions in Africa may not have the ability to collect research data electronically (paper records are still widely used), therefore DMPs need to consider such collection as well as the transcription and electronic transfer of such information. Data to be generated during a research project may require equal attention in a DMP. H3ABioNet tackles these challenges in two ways; 1) by creating awareness of the importance of standardized collection and developing resources that encourage the use and adaption of existing data collection standards and resources in Africa; and 2) by providing "how to guidelines" and training to scientists on various topics including data management, standardization, governance, and more.

Data generation facilities are not always widely available in Africa, and often (particularly so in the case of genomics) research data are generated on other continents which may create issues surrounding IP and data ownership. The DMP thus needs to determine where research data will be generated, as well as how data transfer will occur (both cloud and hard drive solutions are still employed across Africa) and

outline who will be able to access and ultimately own the data. Furthermore, many sensitivities may exist around the collection of African data and these sensitivities must be noted and understood ahead of data generation (e.g. genomic data associated with a stigmatized disease or a vulnerable population) as the data may require increasing levels of protection i.e. increased access control, advanced encryption etc.

The H3Africa consortium consists of multiple sites spread across various African countries, most of which are collecting genomic data associated with specific diseases (i.e. highly sensitive data). The major data types being collected by these projects include genomic sequence data, genotype array files, the associated phenotypes and metadata that is collected along with the samples, and results of any analysis conducted [22]. As these projects are typically funded by the NIH and WT, data are required to be made available to the scientific community through submission to repositories. However, due to sensitivities that exist around this data and due to a general lack of infrastructural capacity on the continent, H3ABioNet was tasked to develop the first formalized human genomic data archive on the continent. This archive assists projects in standardizing their data for submission to public repositories and provides various services such as: cold storage - where projects deposit data and the data remains until researchers have analyzed their data for publication; an encryption/decryption service - for the secure transfer of data to the archive and services to validate, quality control, prepare and submit data to a repository. H3Africa projects are usually requested to appoint a data submitter/manager who registers the project with the archive to alert the archiving team of the project and its expected data types, storage, and privacy requirements. This data manager/submitter then works with the archiving team to ensure all data have been prepared and submitted appropriately and supports the local research team.

7.2. Data management challenges

There remains a lack of research funding on the continent for data management purposes and this is perhaps one of the key difficulties in developing appropriate DMPs on the continent. Moreover, limited human resources, lack of RDM guidance and incoherent policy frameworks, limited formal RDM training, lack of robust and secure technological infrastructure, and limited support and guidance for researchers from the management of the academic institutions on the issue of RDM often hamper the wide implementation of good data management practices in Africa [38,39]. Data storage is still a significant challenge in Africa, especially when it comes to Big Data, and this needs to be carefully considered when drafting a DMP, accounting for storage needs, available infrastructure, and costing requirements.

As mentioned previously, because it is funded by both the NIH and WT, the H3Africa consortium is required to make the data collected and produced as widely available as possible. During deliberations however, the consortium realized that several special considerations would need to be made when sharing the data, due to cultural and ethical issues previously faced by scientists working with African data. Access to the data generated by H3Africa is controlled by a designated Data and Biospecimen Access Committee (DBAC), which needs to consider a number of factors when reviewing an access request. Due to the greater degree of access to software and trained staff in developed countries, publication embargoes are placed on data access, to allow African researchers a sufficient amount of time to exploit their data for their research questions. These researchers also had an opportunity to provide project proposals for data beforehand, to prevent access for duplicate projects in the future. In addition, appropriate acknowledgements of the data need to be provided when data is accessed, and collaboration is encouraged for access requests. Importantly, access is subject to stipulations provided during initial collection of consent. The H3Africa, through H3ABioNet, makes data available in the European Genome-phenome Archive (EGA) and European Nucleotide Archive (ENA) via the H3Africa archiving service [22].

H3Africa have also created a data sharing, access and release policy specific to H3Africa needs which can be accessed here: https://h3africa.org/wp-content/uploads/2020/06/H3Africa-Consortium-Data-Access-Release-Policy-April-2020.pdf.

The key points of the H3Africa data sharing, access and release policy include:

- Maximizing the availability of research data, in a timely and responsible manner.
- Protecting the rights and privacy of human subjects who participated in research studies.
- Recognizing the scientific contribution of researchers who generated the data.
- Considering the nature and ethical aspects of proposed research whilst ensuring the timely release of data.
- Promoting deposition of genomic data in existing community data repositories whenever possible.

8. Recommendations

1. Where possible, data collection should be standardized across the research communities, particularly in biomedical research.
2. African researchers and their graduate students should be encouraged to familiarize themselves with the best practices for drafting effective DMPs for their proposals.
3. Promote and advocate for awareness of RDM within African institutions through academic and research libraries.
4. African librarians need to be trained in RDM with strong support from their institutions.
5. Researchers should create, review, and share DMPs that meet institutional and funder requirements.
6. It is important to design and deliver an online RDM training course including a DMP, which is customized to the African research data management context and requirements.

9. Conclusion

A DMP is a dynamic document and should be reviewed regularly to update and improve it according to the actual needs as the project progresses. A DMP is best prepared at the beginning of the research project; however, it is not too late to start mid-way during the research process – better late than never. A DMP helps to ensure consistent practice in data handling among all project members regardless of future turnover. A DMP should provide the project members, funders and other stakeholders 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.

The paper reviewed the practices and procedures researchers use to collect data sources. For example, to draw from a researcher's prior project or an existing database. There is also a need to determine the media or formats the research wants to collect and use data sources. Managing research data will lead to: 1) meet funding agency requirements; 2) write more competitive grant applications; 3) get credit for your data and increase its impact and visibility; 4) encourage the discovery and use of your data to explore new research questions; 5) improve your data's accuracy, completeness, and usability; 6) ensure long-term preservation of data for future researchers; and 7) comply with ethics and privacy policies.

Even if one is not required by your funding agency, developing a **data management plan** (DMP) at the beginning of a new project will inform good practice throughout the research life cycle. An effective DMP will assist in adhering to the guidelines set by any funding agencies and institutions that are sponsoring the research. It always considers best practice to include budget costs within DMP for data creation, processing, analysis, storage, sharing, and preservation as some Funding Agencies accept these costs in grant applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U24HG006941. H3ABioNet is an initiative of the Human Health and Heredity in Africa Consortium (H3Africa). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] E. Tambo, E.C. Ugwu, J.Y. Ngogang, Need of surveillance response systems to combat Ebola outbreaks and other emerging infectious diseases in African countries, Infect. Dis. Pov. 3 (2014) 29, https://doi.org/10.1186/2049-9957-3-29.
[2] S.A. Tishkoff, F.A. Reed, F.R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A.A. Awomoyi, J.-M. Bodo, O. Doumbo, M. Ibrahim, A.T. Juma, M. J. Kotze, G. Lema, J.H. Moore, H. Mortensen, T.B. Nyambo, S.A. Omar, K. Powell, G.S. Pretorius, M.W. Smith, M.A. Thera, C. Wambebe, J.L. Weber, S.M. Williams, The Genetic Structure and History of Africans and African Americans, Science 324 (5930) (2009) 1035–1044, https://doi.org/10.1126/science:1172257.
[3] R.K. Bains, African variation at Cytochrome P450 genes, Evol. Med. Public Health. 2013 (2013) 118–134, https://doi.org/10.1093/emph/eot010.
[4] E. Matovu, B. Bucheton, J. Chisi, J. Enyaru, C. Hertz-Fowler, M. Koffi, A. Macleod, D. Mumba, I. Sidibe, G. Simo, M. Simuunza, B. Mayosi, R. Ramesar, N. Mulder, S. Ogendo, A.O. Mocumbi, C. Hugo-Hamman, O. Ogah, A. El Sayed, C. Mondo, J. Musuku, M. Engel, J. De Vries, M. Lesosky, G. Shaboodien, H. Cordell, G. Pare, B. Keavney, A. Motala, E. Sobngwi, J.C. Mbanya, B. Hennig, N. Balde, M. Nyirenda, J. Oli, C. Adebamowo, N. Levitt, M. Mayige, S. Kapiga, P. Kaleebu, M. Sandhu, L. Smeeth, M. McCarthy, C. Rotimi, Enabling the genomic revolution in Africa, Science 344 (6190) (2014) 1346–1348, https://doi.org/10.1126/science:1251546.
[5] M.P. Adoga, S.A. Fatumo, S.M. Agwale, H3Africa: a tipping point for a revolution in bioinformatics, genomics and health research in Africa, Source Code Biol. Med. 9 (2014) 10, https://doi.org/10.1186/1751-0473-9-10.
[6] H3ABioNet, Pan African Bioinformatics Network for H3Africa, n.d. http://www.h3abionet.org (accessed March 10, 2021).
[7] W.K. Michener, P.E. Bourne, Ten Simple Rules for Creating a Good Data Management Plan, PLoS Comput. Biol. 11 (10) (2015) e1004525, https://doi.org/10.1371/journal.pcbi.100452510.1371/journal.pcbi.1004525.g001.
[8] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M. E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (1) (2016), https://doi.org/10.1038/sdata.2016.18.
[9] OECD Principles and Guidelines for Access to Research Data from Public Funding, n.d. https://www.oecd.org/sti/inno/38500813.pdf (accessed March 10, 2021).
[10] M. Rossetto, J.-Y. Yap, J. Lemmon, D. Bain, J. Bragg, P. Hogbin, R. Gallagher, S. Rutherford, B. Summerell, T.C. Wilson, A conservation genomics workflow to guide practical management actions, Glob. Ecolohu Conserv. 26 (2021) e01492, https://doi.org/10.1016/j.gecco.2021.e01492.
[11] M. Williams, J. Bagwell, M. Nahm Zozus, Data management plans: the missing perspective, J. Biomed. Inform. 71 (2017) 130–142, https://doi.org/10.1016/j.jbi.2017.05.004.
[12] H. van Zeeland, J. Ringersma, The Development of a Research Data Policy at Wageningen University & Research: Best Practices as a Framework, Liber Quart. 27 (2017) 153–170, https://doi.org/10.18352/lq.10215.
[13] Ricky Erway, OCLC Research, Starting the conversation: university-wide research data management policy, OCLC Research, 2013.
[14] C. Tenopir, S. Talja, W. Horstmann, E. Late, D. Hughes, D. Pollock, B. Schmidt, L. Baird, R.J. Sandusky, S. Allard, Research Data Services in European Academic Research Libraries, Liber Quart. 27 (2017) 23–44, https://doi.org/10.18352/lq.10180.
[15] N. Hall, B. Corey, W. Mann, T. Wilson, Model Language for Research Data Management Policies, n.d. https://www.fosteropenscience.eu/content/model-language-research-data-management-policies (accessed March 10, 2020).
[16] ANDS, Outline of a Research Data Management Policy for Australian Universities/Institutions, n.d. http://nlwra.gov.au/files/pages/2613/module-2-data-management-.
[17] LEARN, LEARN Toolkit of Best Practice for Research Data Management, 2017. https://doi.org/10.14324/000.learn.00.
[18] LEARN, SURVEY: Is your institution ready for managing research data?, 2017. http://www.learn-rdm.eu/material/leru_roadmap_for_research_data.
[19] T. Miksa, P. Walk, P. Neish, RDA DMP Common Standard for Machine-actionable Data Management Plans, 2020. https://doi.org/10.15497/rda00039.
[20] J. Cardos, D. Proença, J. Borbinha, Machine-Actionable Data Management Plans: A Knowledge Retrieval Approach to Automate the Assessment of Funders' Requirements, in: J. Jose, et al. (Eds.) Advances in Information Retrieval. ECIR (2020) Lecture Notes in Computer Science, vol. 12036, Springer, Cham, 2020. https://doi.org/10.1007/978-3-030-45442-5_15.
[21] S. Simms, S. Jones, D. Mietchen, T. Miksa, Machine-actionable data management plans (maDMPs), Res. Ideas Outc. 3 (2017) 3e13086, https://doi.org/10.3897/rio.3.e13086.
[22] Z. Parker, S. Maslamoney, A. Meintjes, G. Botha, S. Panji, S. Hazelhurst, N. Mulder, Building Infrastructure for African Human Genomic Data Management, Data Sci. J. 18 (2019), https://doi.org/10.5334/dsj-2019-047.
[23] Nicola J. Mulder, Ezekiel Adebiyi, Marion Adebiyi, Seun Adeyemi, Azza Ahmed, Rehab Ahmed, Bola Akanle, Mohamed Alibi, Don L. Armstrong, Shaun Aron, Efejiro Ashano, Shakuntala Baichoo, Alia Benkahla, David K. Brown, Emile R. Chimusa, Faisal M. Fadlelmola, Dare Falola, Segun Fatumo, Kais Ghedira, Amel Ghouila, Scott Hazelhurst, Itunuoluwa Isewon, Segun Jung, Samar Kamal Kassim, Jonathan K. Kayondo, Mamana Mbiyavanga, Ayton Meintjes, Somia Mohammed, Abayomi Mosaku, Ahmed Moussa, Mustafa Muhammd, Zahra Mungloo-Dilmohamud, Oyekanmi Nashiru, Trust Odia, Adaobi Okafor, Olaleye Oladipo, Victor Osamor, Jellili Oyelade, Khalid Sadki, Samson Pandam Salifu, Jumoke Soyemi, Sumir Panji, Fouzia Radouani, Oussama Souiai, Özlem Tastan Bishop, Development of Bioinformatics Infrastructure for Genomics Research, Global, Heart 12 (2) (2017) 91, https://doi.org/10.1016/j.gheart.2017.01.005.
[24] M. Williams, J. Bagwell, Zozus M. Nahm, Data management plans: the missing perspective, J. Biomed. Inform. 71 (2017) 130–142, https://doi.org/10.1016/j.jbi.2017.05.004. S1532-0464(17)30099-0.
[25] A. Lefebvre, B. Bakhtiari, M. Spruit, Exploring research data management planning challenges in practice, Inform. Technol. 62 (1) (2020) 29–37, https://doi.org/10.1515/itit-2019-0029.
[26] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, M. Vingron, Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, Nat. Genet. 29 (4) (2021) 365–371, https://doi.org/10.1038/ng1201-365.
[27] N.R. Anderson, E.S. Lee, J.S. Brockenbrough, M.E. Minie, S. Fuller, J. Brinkley, P. Tarczy-Hornoch, Issues in biomedical research data management and analysis: needs and barriers, J. Am. Med. Inform. Assoc. 14 (4) (2007) 478–488, https://doi.org/10.1197/jamia.M2114.
[28] The Digital Curation Centre, Overview of funders' data policies. https://www.dcc.ac.uk/guidance/policy/overview-funders-data-policies (accessed March 10, 2021).
[29] NIH Data Sharing Policy and Implementation Guidance, n.d. https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#time (accessed March 11, 2021).
[30] Medical Research Council (MRC), Section 2 Guidelines and standards A Planning and conducting MRC-funded research, n.d., pp. 7–21. https://mrc.ukri.org/documents/pdf/good-research-practice-guidelines-and-standards/.
[31] Medical Research Council (MRC), Sharing research data to improve public health: full joint statement by funders of health research, n.d. https://wellcome.org/what-we-do/our-work/sharing-research-data-improve-public-health-full-joint-statement-funders-health (accessed March 10, 2021).
[32] Tomasz Miksa, Stephanie Simms, Daniel Mietchen, Sarah Jones, Francis Ouellette, Ten principles for machine-actionable data management plans, PLoS Comput. Biol. 15 (3) (2019) e1006750, https://doi.org/10.1371/journal.pcbi.100675010.1371/journal.pcbi.1006750.g00110.1371/journal.pcbi.1006750.g00210.1371/journal.pcbi.1006750.g003.
[33] R. Otsuji, S. Labou, R. Johnson, G. Castelao, B. Villas Boas, A.-L. Lamprecht, C. Martinez Ortiz, C. Erdmann, L. Garcia, M. Kuzak, P. Andrea Martinez, L. Stokes, N. Simons, T. Honeyman, S. Wise, J. Quan, S. Peterson, A. Neeser, L. Karvovskaya, O. Lange, I. Witkowska, J. Flores, F. Bradley, K. Hettne, P. Verhaar, B. Companjen, L. Sesink, F. Schoots, E. Schultes, R. Kaliyaperumal, E. Toth-Czifra, R. de Miranda Azevedo, S. Muurling, J. Brown, J. Chan, L. Federer, D. Joubert, A. Dillman, K. Wilkins, I. Chandramouliswaran, V. Navale, S. Wright, S. di Giorgio, A. Mandela Fasemore, K. Förstner, T. Sauerwein, E. Seidlmayer, I. Zeitlin, S. Bacon, K. Hannan, R. Ferrers, K. Russell, D. Whitmore, T. Dennis Organisations, Top 10 FAIR Data & Software Things Sprinters, 2019. https://doi.org/10.5281/zenodo.3409968.
[34] G.T. Chiang, P. Clapham, G. Qi, K. Sale, G. Coates, Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute, BMC Bioinf. 12 (2011) 361, https://doi.org/10.1186/1471-2105-12-361.
[35] P.A. Harris, R. Taylor, J. Thielke, N. Payne, J.G. Conde Gonzalez, Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support, J Biomed. Inform. 42 (2) (2009) 377–438, https://doi.org/10.1016/j.jbi.2008.08.010.
[36] T. Davey, D. Fripp, J. Kaye, Jisc Open Research Hub – Supporting Open Scholarship, in: The 14th International Conference on Open Repositories, Hamburg, Germany, Zenodo, 2019. https://doi.org/10.5281/zenodo.3554342.
[37] M.C. Schatz, A.A. Philippakis, E. Afgan, E. Banks, V.J. Carey, R.J. Carroll, A. Culotti, K. Ellrott, J. Goecks, R.L. Grossman, I.M. Hall, K.D. Hansen, J. Lawson, J.T. Leek, A. O'Donnell Luria, S. Mosher, M. Morgan, A. Nekrutenko, B.D. O'Connor, K. Osborn, B. Paten, C. Patterson, F.J. Tan, C. Overby Taylor, J. Vessio, L. Waldron, T.

Wang, K. Wuichet, AnVIL Team, Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL), bioRxiv (2021) 436044. https://doi.org/10.1101/2021.04.22.436044.

[38] B.K. Avuglah, P.G. Underwood, Research Data Management (RDM) Capabilities at the University of Ghana, Legon, Libr. Philosop. Pract. (e-Journal) (2019) 2258. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=5670&context=libphilprac.

[39] Josiline Chigwada, Blessing Chiparausha, Justice Kasiroori, Research Data Management in Research Institutions in Zimbabwe, Data Sci. J. 16 (0) (2017) 31, https://doi.org/10.5334/dsj-2017-03110.5334/dsj-2017-031.s1.