Version: September 2017

RESEARCH DATA MANAGEMENT
Good Practice Note

Prepared by: CGIAR Internal Audit Unit

Table of contents

FOREWORD

What is a GPN

A Good Practice Note (GPN) is a document themed around a specific risk or control-related area.
It is developed by the CGIAR IAU with contributions of subject-matter specialists, leveraging knowledge accumulated within the CGIAR System and reflecting good practices suggested by professional bodies or standard setters, and implemented by Centers and/or other external organizations.

GPNs aim to summarize, circulate and promote existing knowledge around the System and can be used to benchmark existing arrangements against good practices and to improve knowledge, processes and operations at Center and System levels.

What it is not

GPNs are not and should not be interpreted as minimum standards, policies, guidelines or requirements, as practices mentioned in the GPN may not be relevant to or applicable in all Centers.

1. INTRODUCTION

The core business of CGIAR Centers is to deliver agricultural research for development. Centers do it by conducting activities organized in projects that can include scientific research as well as scaling out activities. Such projects are initiated, conducted and implemented in the context of the CGIAR Strategy and Results Framework (SRF). Defined collectively as "CGIAR Research", they encompass the Portfolio of CGIAR Research Programs (CRPs), Platforms and other projects relevant to the SRF.
As an outcome of these research activities, a large amount of valuable data and information products are generated. These results are classified as international public goods and CGIAR is committed to the dissemination and use of these outputs to achieve maximum impact. To ensure this is achieved, effort and focus must be put in the establishment of good practices in the management of research data and related information products.
Within the CGIAR system, each Center defines its research data management approach in alignment with the common approach provided by the CGIAR through its "*Open Access and Data Management Policy".*

The purpose of this GPN is to:

- provide reference materials to Center management and internal audit teams as to existing good practices in research data management
- raise awareness of the good practices available and support self-assessment against them
- provide suggestions on adaptation and adoption of the good practices in research data management at a Centre level
- support improvements in the research data management processes and controls.

1.1 What is Research Data?

In this document, "research data" is defined as factual records (data sets, surveys, publications, images, video etc.) used as primary sources for scientific research and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated. Research data can either be digital i.e. computer readable or in a physical format such as laboratory notebooks, laboratory specimens etc.

1.2 Benefits of research data management

Good research data management is not a goal in itself, but rather the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and re-use. Effective data management will support FAIR[1] data principles i.e. data will be:

A) **Findable**
B) **Interoperable**
C) **Accessible**

[1] An outcome of a workshop held in Leiden, Netherlands, in 2014, named 'Jointly Designing a Data Fairport', that brought together a wide group of academic and private stakeholders who had an interest in overcoming data discovery and reuse obstacles.

D) **Reusable**

The principles were developed to:

- Encourage systematic documentation and descriptions of the research data
- Provide guidelines and procedures ensuring consistency
- Safeguard against data loss
- Ensure confidentiality and ethical compliance
- Comply with intellectual property rights such as copyright
- Allow researchers to validate and verify published results
- Enable collaborative research opportunities thereby increasing the potential scale and scope of research
- Prevent duplication of research within a specific field
- Allow data sharing and future use when the data is preserved in retrievable formats
- Increase citations for a researcher
- Allow for data replication or reproducibility
- Increase the accuracy or reliability of data
- Ensure research data integrity
- Help researchers to meet requirements of funders
- Meet the requirements of publishers who are increasingly requiring submission of underlying data for published work to support the research
- Improve management of data and the research process.

1.3 Potential risks associated with poor research data management

Effective research data management also plays a vital role in managing research risk. All research is subject to a range of data-related risks such as data loss or corruption, and privacy or copyright breaches. These risks come with significant, potentially catastrophic impacts. Effective research data management can go a long way towards preventing and managing such risks.

1.3.1 Risks related to Research Data Management, their sources and implications

RISKS	SOURCES	IMPLICATIONS
• Data is not secure • Poor quality research data/metadata • Segmented data • Unwarranted disclosure of restricted research data • Breach of funder's data management obligations • Breach of IP rights • Breach of ethical research protocols	• Poor management of access rights allowing unauthorized access to data • Unclarity of roles and responsibilities • Insufficient technological support for data management • Lack of standardization • Lack of formal data management processes • Lack of resources and/or expertise to establish data management processes • Lack of training in data management • Scientific fraud • Poor coordination of organization's data infrastructure • Lack of oversight of data management • Lack of understanding of funders/legal requirements	• Loss of data integrity • Loss of valuable research data • Data is not accessible • Breach of confidentiality requirements • Data cannot be re-used • Loss of research credibility • Lack of interoperability of data • Duplication of efforts • Loss of funding • Reputational damage • Danger to public health • Misalignment of resources

2. State of RDM at CGIAR

2.1 CGIAR System requirements

2.1.1 CGIAR Open Access and Data Management Policy

At the CGIAR level, expectations on research data management have been established through the formal approval of the *CGIAR Open Access and Data Management Policy* on 02 October 2013 http://library.cgiar.org/bitstream/handle/10947/4488/Open%20Access%20Data%20Management%20Policy.pdf?sequence=1 . Implementation of and compliance with this policy by the CGIAR System office, its members and their partners within the scope of the Strategy and Results Framework ("SRF") and the CGIAR Research Programs ("CRPs") would be phased over a transition period from the effective date of

the policy for an initial period of 5 years, with comprehensive implementation by the end of 2018. The policy is to be utilized in conjunction with the *CGIAR Open Access and Data Management Implementation Guidelines* http://www.cgiar.org/?s=CGIAR+Open+Access+and+Data+Management+Implementation+Guidelines.+&s_area=all .

The implementation of the CGIAR OA/DM policy is currently incorporated in the "Organize" module of Big Data platform which aims to "to harness the capabilities of Big Data to accelerate and enhance the impact of international agricultural research" http://www.cgiar.org/about-us/our-programs/cgiar-platform-for-big-data-in-agriculture-2017-2022/ .

2.1.2 CGIAR Principles on the Management of Intellectual Assets

This policy was effective as part of the Common Operational Framework as of 7 March 2012 and was approved by the Consortium Board on 1 March 2012 and by the Fund Council on 7 March 2012. It can be found here http://www.cgiar.org/consortium-news/principles-on-management-of-intellectual-assets-approved/ . The policy documents the principles agreed by the Consortium (the predecessor of the System Organization) and the Fund Council (now System Council) with respect to the management of Intellectual Assets produced or acquired by the Consortium (now System Organization) and/or the Centers.

2.2 RDM maturity model

IAU has developed an RDM maturity model that can be used by management to understand the level of development of the data management frameworks in a Center. The matrix is described below and is based on good practice described throughout the document. The highlighted areas reflect on how the majority of the Centers measure against the RDM maturity continuum based on four RDM audit reviews carried out by IAU in 2015-2016.

The matrix can be used to assess the current state of RDM activities and systems at a Center as well as to identify and plan improvement programs. Any feedback, comments or questions about the maturity model should be addressed to CGIAR IAU team.

CGIAR
Science for a food-secure future

RESEARCH DATA MANAGEMENT MATURITY MODEL

Maturity Level	Level 1	Level 2	Level 3	Level 4	Level 5	Observations
Key characteristics	**Process is disorganized & adhoc**	**Process is under development**	**Process is standardized , communicated**	**Process is managed , measured**	**Focus is on continuous improvement**	**Observations from recent audits of Centers**
Institutional policies and procedures	Policies and procedures may be underdeveloped, not up to date and/or inconsistent	Policies and procedures are developed and harmonized	Policies and procedures are promulgated and absorbed into behaviors	Policies and procedures accepted as part of the culture and subject to audit	Policies and procedures subject to review and improvement	RDM Policies still in draft status and not yet fully operational
IT infrastructure	IT infrastructure provision is patchy, disorganized	• Funds are invested in technology and skills. • Responsibilities are defined, processes are established, defined and documented	• Management shows active support. • Facilities are well defined and communicated, standardized and integrated	• Funding adapted to need. • Management actively engaged	• Concerted effort to maintain, update and publicize infrastructure • Feedback used to optimize services	Lack of proper definition of roles and responsibility between IT teams and data management teams leading to conflicts, inefficiency, wastage of resources and duplication
Repositories	No institutional repositories available. Documents are on departmental file	• Unclear policies on acquisition and maintenance of repositories, too	• Standardization in use of repositories across similar research areas	• All valuable research data stored and managed in	Repositories kept continually up to date with changes in metadata standards	• Investment in institutional repositories still in infancy

Maturity Level	Level 1	Level 2	Level 3	Level 4	Level 5	Observations
Key characteristics	**Process is disorganized & adhoc**	**Process is under development**	**Process is standardized , communicated**	**Process is managed , measured**	**Focus is on continuous improvement**	**Observations from recent audits of Centers**
	systems, user desktops, google docs and other places	many adhoc repositories in use • Institutional repositories available are not compliant with the CG Core metadata standards	• Clear policies the on acquisition and maintenance of repositories that discourage proliferation (e.g. for individual projects)	institutional repositories • Repositories are FAIR and CG Core compliant and can provide metrics for usage monitoring • Institutional inventory available of all repositories and their maintenance	and technological advancements	• Too many unregulated repositories and data storage solutions with no clarity on ownership and accountability of repositories once project ends
Training	No training programs are available for researchers	• Training is ad-hoc and inconsistent • Participation and attendance is low • Trainings focused mainly on HQ	• Institution consciously invests in trainings and development of reference materials, support packs (e.g. CGIAR OA/OD support pack)	Active participation in training and widespread availability of training for all offices	Researchers feedback used extensively to update and improve trainings	• Lack of sufficient training programs for new research staff, interns and visiting scientists • Training only focused on HQ staff

Maturity Level	Level 1	Level 2	Level 3	Level 4	Level 5	Observations
Key characteristics	**Process is disorganized & adhoc**	**Process is under development**	**Process is standardized , communicated**	**Process is managed , measured**	**Focus is on continuous improvement**	**Observations from recent audits of Centers**
			• Person/ Unit responsible for RDM trainings ideally identified and working within a RDM support team			
RDM support team or services	No institutional unit/ support services team on RDM matters	• Dedicated RDM team available with limited resources • no clear strategic direction, role & responsibilities	• Dedicated RDM team/unit with well-defined roles and responsibilities • services inconsistent and poorly publicized	• Widespread take up of RDM services • Consistent, demand/need driven services	Continuous interaction with researchers and improvements of support services based on researcher feedback and needs	• Lack of institutional RDM support services team • RDM support services with limited resources i.e. budget and staff
Risk Management	No identification or assessment of risks related to RDM	Risks identified are patchy and generic, no proper risk assessment undertaken	Risks related to RDM appropriately identified and mitigating plans put in place	Risk information and mitigating plans are shared with managers and staff to help identify responsibilities and respond appropriately to the risks	Risks related to RDM are regularly reviewed, updated and communicated as appropriate by the organization	• Weak coverage of RDM risks as part of the Institutional risk management process • Lack of clear accountability over RDM risks
Awareness	Limited awareness of the purpose or	Usually senior management is aware of existence	Management understands how RDM	Management understands the long-term RDM	Both management and staff actively	• Awareness programs

Maturity Level	Level 1	Level 2	Level 3	Level 4	Level 5	Observations
Key characteristics	**Process is disorganized & adhoc**	**Process is under development**	**Process is standardized , communicated**	**Process is managed , measured**	**Focus is on continuous improvement**	**Observations from recent audits of Centers**
	value of Research Data Management	of RDM initiatives but don't actively promote or support it.	benefits/impacts the organization and actively promotes awareness such that all staff are aware of the RDM initiative	strategy and takes part in it. All staff understand the purpose of RDM and their role in it.	promote and support RDM	concentrated on HQ staff • Lack of ownership of awareness campaigns and programs between RDM support staff or HR learning and development units
Data Management Plans (DMPs)	• DMPs are not a priority to the institution • No DMP related polices or guidelines	DMPs are adhoc, inconsistent and normally undertaken only where it is part of a funder proposal requirement	• Standard DMP tools and templates used or customized by the institution (see CGIAR OA/OD support pack for template) • Clear policies and guidelines on DMP	DMPs are consistency incorporated as part of the project lifecycle for all projects	Continuous monitoring of DMP implementation and improvements	DMP not consistently used
Metadata	• Metadata is not a priority to the institution. Minimal, if any,	Metadata has been defined as important to the institution. Work is underway to	Metadata best practices are produced at the dataset and repository levels	• Metadata collection/validation responsibilities assigned to	• Metadata management is a top priority, and is used to document all data.	Lack of defined institutional policies and standards on metadata

Maturity Level	Level 1	Level 2	Level 3	Level 4	Level 5	Observations
Key characteristics	Process is disorganized & adhoc	Process is under development	Process is standardized , communicated	Process is managed , measured	Focus is on continuous improvement	Observations from recent audits of Centers
	metadata is maintained. • Limited understanding of types and value of metadata. • No Metadata related policies	implement a process to use standard metadata and datasets. Work is underway to implement or map to the CG Core at the repository level.	and made available but best practices are focused on the metadata associated only with structured data. Repository metadata schema has been implemented and mapped to CG Core.	named individuals for each projects. • Policies requiring the regular auditing of metadata in specified systems are adopted as part of official Center data policies • Metadata implementation at both repository and dataset levels (where appropriate) is enforced.	• A dedicated metadata focal point exists within RBM support team for e.g. strategically advance metadata capabilities and more effectively leverage existing metadata. • Metadata policy covers both structured and unstructured (non-tabular) data and is enforced	definition and documentation
Documentation of RDM process, SOPs	• Documents are on departmental file systems, user desktops, google docs and other places in	• Some work has already been accomplished to move the electronic records off of user desktops and onto	SOPs available e.g. organization of data, version control file formats etc.	• SOPs consistently implemented across projects at HO and field offices.	Continuous monitoring and improvement of SOPs	Lack of defined procedures and toolkits to assist researchers with best practice on documentation

Maturity Level	Level 1	Level 2	Level 3	Level 4	Level 5	Observations
Key characteristics	**Process is disorganized & adhoc**	**Process is under development**	**Process is standardized , communicated**	**Process is managed , measured**	**Focus is on continuous improvement**	**Observations from recent audits of Centers**
	unstructured way. • Little or no version control and no focus on data formats used	secured storage areas. • Efforts in place to standardize documentation practices		• Regular auditing		

3. ROLES AND RESPONSIBILITIES TO ENSURE GOOD RDM

Data management is not just the responsibility of a researcher who has created or collected the data. Various parties are involved in the research process and may play a role in collecting data, safeguarding it and facilitating data sharing. It is therefore crucial that roles and responsibilities are assigned and not just presumed. For collaborative research, assigning roles and responsibilities across partners is important. In this GPN, we have highlighted what we consider are the key roles and responsibilities of the various parties involved in the research project data cycle.

A) **Center Management**

- Ratifies policies that articulate the core RDM principles and acts as a framework for guidelines and service design.
- Provides safe, secure and sustainable funding for infrastructure and support services to make data and information available respecting the rights of stakeholders in terms of confidentiality, intellectual property and data ownership.
- Establishes a representative, balanced and appropriately equipped steering/working group that will reflect the interests of essential stakeholders.
- Ensures that performance assessment covers data and information management. This should include incentives, recognition and rewards for making data and information more open and accessible.

B) **Research Data and Information Management steering committee / working group**

- Provides guidance and recommendations for strategic directions and priorities
- Oversees the implementation of the Data and Information Management Framework
- Approves and endorses any data management policies, processes, standards and guidelines
- Advises on the higher-level strategic issues that must be addressed during RDM framework design.

C) **Data Management support teams**

The support teams that will deliver RDM services may typically be categorized as the library, information technology, records management, sharing and dissemination, and research administration functions, although this list is not exclusive. These teams should therefore include expertise across data and information management, legal/IP, and communications management and liaise with other units e.g. project management. It is important to recognize that in many institutions these groups will not have previously worked as a cohesive unit or partnership. To deliver RDM services they will together provide the effort to:

- Undertake the actions defined by the steering group
- As traditionally independent units, reorganize into a partnership to deliver a seamless RDM service.
- Develop policies, procedures, toolkits and guidelines for research data management
- Identify the nature of the institution's data assets and manage institutional data management infrastructure for adequate data management
- Develop and implement proposals, plans and budgets for the technological and human infrastructures necessary to deliver RDM services

- Provide a range of capacity development programs on data management and facilitate training opportunities for project managers and researchers
- plan and undertake a program of advocacy to promote the key aspects of effective research data management, explaining in universally accessible terms its obligations, benefits and the services anticipated.

D) **Directors of Research, Heads of Research Units and Research Project Managers**

- Have overall accountability for research data management (RDM) and for ensuring that it is implemented
- Set the RDM culture and practice (based on good practice guidance) and delegate specific responsibilities as appropriate
- Make decisions (after consultation within the team) about issues such as data access, data sharing, long-term retention of data
- Monitor RDM practices and ensure RDM requirements are met
- Allocate the necessary resources to conduct effective RDM to ensure that appropriate expertise exists e.g. in budgeting FTEs across projects and in domains such as IT, legal, data management etc.
- Ensure that staff are aware of their responsibilities and obligations in effective management of data and information management and identify or promote continuous training where gaps in these skills are identified
- Promote and incentivize good practice in data and information management.

E) **Researchers / Scientists**

As the creators and users of research data, researcher engagement is crucial in the development of RDM practices. Without their active involvement and support the success of an RDM practices is bound to be limited. Whilst management will define expectations and support staff will deliver services, it is the responsibility of researchers to:

- Include appropriate consideration of the cost and time implications of data storage and management within grant proposals (via part-time FTEs as appropriate)
- Ensure that all relevant research data are assessed for adherence to FAIR in an interoperable repository, unless specified otherwise in the data management plan
- Develop and record appropriate procedures for the collection, annotation, storage and back-up, QA/QC, use, re-use, access, and retention of the research data associated with research programs
- Document agreements for research data management when involved in a joint research project, collaborative research backed by appropriate contractual agreements and considering consequences of privacy protection laws
- Ensure that the integrity and security of data is maintained
- Work with the data management unit to ensure that good practice in research data management is consistently followed
- Ensure researcher views are represented by contributing to steering/working groups
- Clearly articulate the particular requirements, opportunities and obstacles encountered within research
- Collaborate in the gathering of requirements and the testing of solutions and methods proposed

- Champion the adoption of approved methods and services within their research teams.

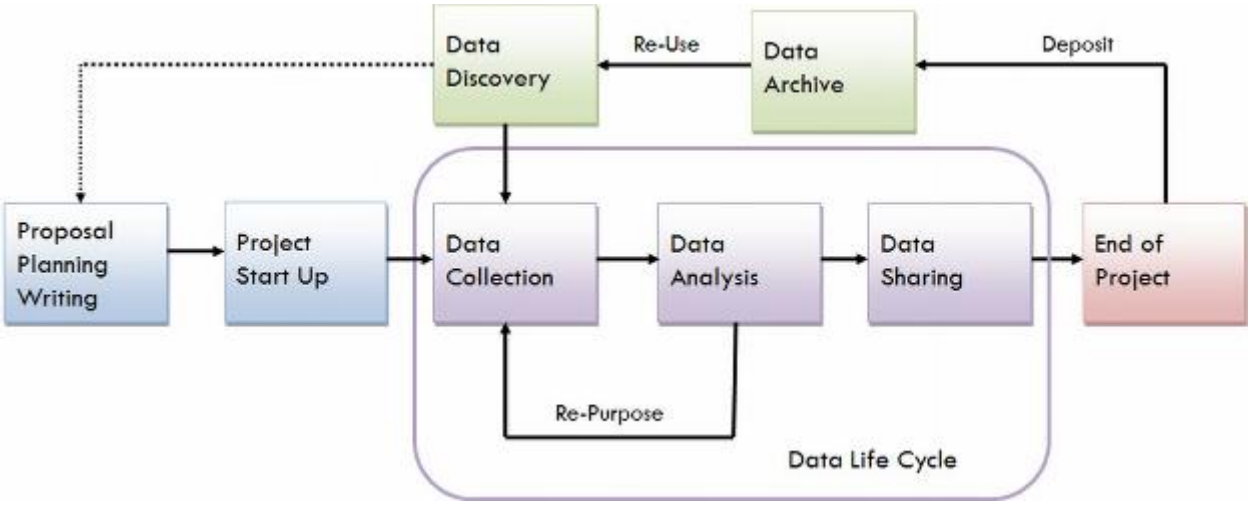
4. RECOMMENDED PRACTICE

Centers increasingly realize that research data is a core asset and is indeed vital to the very existence and relevance of these organizations. It is therefore important that appropriate focus and effort be put in place towards ensuring that this core asset is well managed. There are no single formulae towards how this can be achieved. However, use of a structured approach will help to accomplish that. In this section, we put forward key elements of good management of research data.

4.1 Research Data Life Cycle

Data often have a longer lifespan than a research project that creates them. Researchers may continue to work on data after funding has ceased, follow-up projects may analyze or add to the data, and data may be re-used by other researchers. Therefore, well organized, well documented, preserved and shared data are invaluable to advance scientific inquiry and to increase opportunities for learning and innovation.

Below is a chart describing data life cycle within a project and beyond[2]



At each stage of the data life cycle, specific activities need to be implemented.

4.1.1 Proposal planning & writing

Over the past few years, funders have begun to realize and recognize that planning for data management before research is critical to the project and therefore they have increasingly required this to be included

[2] Courtesy of the University of Virginia, USA https://data.library.virginia.edu/data-management/

as part of the proposal submission process. If data management is not considered during the project proposal development stage, it is unlikely that time and resources would be available to meet funders' and other stakeholders' expectations of a given project.

According to Article 4.1.9 of the Open Access and Data Management Policy: "Open Access and Data Management Plans should be prepared in order to ensure implementation of this Policy. Such Plans shall, in particular, outline a strategy for maximizing opportunities to make information products Open Access." In keeping with the FAIR principles, the policy further notes the importance of ensuring both the syntactic and semantic interoperability of data. At a Center level, therefore, guidance or templates that outline what is expected in a good data management plan should be developed. However, it is the responsibility of each project manager to ensure that their project has a defined data management plan aligned with the goals of the OADM policy and supported by a solid budget.

There are some basic components that need to be included in data management plans:

- A description of the types of data that will be collected or generated during the project
- The standards (metadata, interoperability) that will be used for those data and their associated metadata
- Plans for archiving, preserving and making accessible of the data generated
- Plans for discussing data quality
- Reuse conditions (e.g. licensing, crediting etc.)
- A description of the resources needed to accomplish data management, including personnel hardware, software and budgetary requirements.

A suggested example of a data management plan checklist from Digital Curation Center DCC can be found at http://www.dcc.ac.uk/resources/data-management-plans/checklist .

4.1.2 Project start up

Everyone involved in data collection, analysis, re-use and storage should follow SOPs, understand the data filing, file naming and back up protocols. When folders are used, they should follow an agreed standard where each project, laboratory experiment or sample group is logically placed in a hierarchical order. Researchers need to adhere to disciplinary standards and maintain consistency in file naming and version control to ensure ease of re-use for all. Below is a table that provides an example of a file naming guide.

Naming Standard	Description of Naming Standard
Numbering Standards	Specification of digit numbers to ensure a consecutive listing of files
Date Standards	Specification of date formats to ensure a consecutive listing of files e.g. 'YYYY-MM-DD'
Punctuation Standards	No punctuation or spaces except for underscores to partition words. The period sign should only precede the file extension e.g. 'project_101_sample_001.xls'

Vocabulary Standards	Maintain disciplinary standards in vocabulary, language and abbreviations e.g. ‘project_01_sample_pcr_001’ or ‘project_101_sample_microarray_101’
File Version Numbering	Label the file versions in numerical terms e.g. '1.0, 1.1, 1.2 etc.'
File Version Description	Complete file naming appropriately through the use of descriptive terms at the end of the document name e.g. 'draft_1, draft_2, final_1 etc.'

A) **Version control**

Version control facilitates data quality controls during a project where constant re-drafting and revision is occurring by numerous researchers. Mechanisms must be put into place to distinguish between the different versions. Version control management can be achieved through:

- research data access and editing privilege control;
- selecting one individual to handle all manual editing of data; and
- the use of software.

B) **Metadata standards**

Metadata is data that describes data. It is the documentation that accompanies the research data which makes it discoverable and usable over time. Metadata helps facilitate:

- discoverability of data
- data identification
- data association with publications and related datasets; and
- quality assurance and validation of data.

According to the CGIAR Open Access and Data Management implementation guideline of 2014 *“The CG Core Metadata Schema (CG Core) would be a common framework for CGIAR Consortium Member Centers, CRPs, and other entities to present and share metadata in consistent ways across the network of CGIAR repositories. CG Core is based on Dublin Core, a widely-used metadata standard, with a limited number of additional elements specific to the CGIAR environment.”*

The benefits of having such a standardized metadata schema is to ensure that data is interoperable discoverable for re-use by the wider community across repositories. Having defined metadata standards that provide guidance for Center scientists ensures that there is consistency across all research areas in the way data is documented and re-used. Any non-CG Core schema used should be mapped to the CG Core to allow resource discovery via the CGIAR Big Data Platform.

Metadata standards exist to provide standardized descriptions such as Dublin Core. It is a requirement intended to establish a common understanding of the meaning of the data, to ensure correct and proper use and interpretation of the data by its owners and users. It is critical that a Center has formally prescribed metadata specifications that it adheres to.

The table below provides examples of discipline specific Metadata standards:

Discipline	Metadata Standard
Humanities Data	• The Text Encoding Initiative • The Visual Resources Association Core • Dublin Core • Functional Requirements for Bibliographic Records (FRBR)
Geospatial Data	• The Content Standard for Digital Geospatial Metadata (CSDGM) • ISO Standard for Geographic Information (ISO 19115:2003) • ANZLIC
Social sciences Data	• Data Documentation Initiative (DDI)
Scientific Data	• CCLRC Scientific Data Model
Multimedia	• NISO Z39.87-2002 Technical Metadata for Digital Still Images • MPEG-7

C) **Data formats**

Research data should be secure and retrievable for long term use. During data collection and analyses, researchers may select specific data formats. Conversion of data into standard interchangeable formats may be necessary for preservation purposes. As future access and re-use of data may be affected by proprietary formats, it is advisable to use open formats.

When selecting file formats consideration must be given to:

- Data collection, analysis and sharing
- Discipline-related standards
- Software and hardware compatibility and longevity during the data retention period, and
- Preference of proprietary software as opposed to open source software. A list of open formats for various types of data can be found on https://en.wikipedia.org/wiki/List_of_open_formats.

4.1.3 Data collection, analysis and sharing

4.1.3.1 Data Collection

Data collection refers not only to what information is recorded and how it is the recorded, but also to how a particular research project is designed. Although information necessary to data collection methodology varies by project, the aim of successful data collection should always be to uphold the integrity of the project, the institution, and the researchers involved.

Data collection may seem tedious or repetitive, but the data produced must be reliable and valid to ultimately prove or disprove hypotheses and justify or counter a body of research. In addition, thorough data collection accomplishes the following:

- Allows independent researchers to replicate the process and evaluate results

- Impresses upon research team members the importance of data management
- Validates the rationale behind a research project
- Provides justification to sponsors for expenditures and project decisions
- Yields reliable and valid results, and hypothesis testing.

Data collection should ensure that data is:

- Reliable. Data collection guidelines and methodologies should be carefully developed before the research begins. The researchers must determine what sort of data will be collected and how this data will be analyzed. For data to be considered reliable, data collection should occur consistently and systematically throughout the course of the project, ideally via digital tools where possible
- Valid. Collecting valid data ensures that when research is evaluated it will be deemed good science meaning that the research is both precise and honest. Thorough data collection should thus include a continuous system for rigorously evaluating quality, assessing effective or deficient elements in the project protocol or the research team's techniques.

4.1.3.2 Data repository

Per Article 4.1.2 of the CGIAR Open Access and Data Management Policy: “Stable, permanent, Open Access repositories shall be utilized, to enable users and other sites and search engines to access or locate information products, including application programming interfaces (APIs) or other mechanisms enabling those information products to be available from the CGIAR website and associated web-based products. Preference should be given to existing repositories to minimize the number of repositories in use (and the interoperability challenges presented by multiple incidences of repositories.)”

Individual and project level solutions such as folders in personal machines, do not satisfy the goals of re-use and sharing of data. The reason they remain so widely used is their ease, low cost and that they can be set up rapidly to answer some of the research needs during a research project. To steer researchers away from such quick fix solutions, institutions need to encourage researchers to plan for how they will store data and budget for it at the beginning of a project as donors are also increasingly expecting. For institutional repositories which adhere to the CG Core metadata should be the solution of choice.

4.1.3.3 Ethics, privacy, consent and legal issues

Centers are expected to maintain high ethical standards while conducting their research. A lapse in this area may have detrimental impact to the center both from a legal as well as reputational perspective. It is therefore critical that the Centers ensure that data collection and sharing is undertaken in an ethical manner.

To appropriately manage these risks, a Center can consider the following:

A) **Obtaining of appropriate consent and anonymization of sensitive information**

- Researchers are expected to obtain informed consent from data providers when collecting primary data from people for the use of the information collected. To ensure that consent is informed, consent must be freely given with sufficient information provided on all aspects of participation and data use.

Researchers should:

- inform participants on how research data will be stored, preserved and used in the long-term
- inform participants on how confidentiality will be maintained, e.g. by anonymizing data
- obtain informed consent, either written or verbal, for data sharing.

- Personal data should not be disclosed from research information, unless a respondent has given specific consent to do so. Thus, before data obtained from research with people can be published or shared with other researchers, it needs to be anonymized so that individuals, organizations and businesses cannot be identified from the data.

B) **Establishing legal ownership of the research data**

It is imperative that the ownership of research data is clarified prior to the commencement of a project. Future storage and re-use are directly affected by the intellectual property rights of research data.

Ownership of research data may be influenced by:

- the commercial potential of the research data
- whether the research data is acquired through organizational collaborations
- project funding agreements; and
- whether third-party data has been utilized during the conduct of research.

Intellectual Property Rights (IPRs) permit researchers to control the use of their research data. IPRs such as copyright could ensue automatically as data is being accrued whereas others may require patents depending on the research.

The ownership and management of intellectual property at the Center is governed by the Center's institutional Intellectual Property Policy and the CGIAR Principles on the Management of Intellectual Assets.

Acknowledgement to the adherence to these polices, the relevant clauses on intellectual property should be included as part of staff employment contracts, partnerships agreements as well as the general staff code of conduct.

C) **Adherence to privacy laws and regulations**

Depending on the location of the Center, there are various privacy laws and regulation that may be specific to that country in which the Center operates. It is therefore critical that these be considered while outlining the data sharing protocols.

For Centers within the commonwealth, the Privacy Act 1988 requires compliance with the Information Privacy Principles (IPPs) regarding personal information.

Other legislation that may impact on sharing of confidential information include:

- Data Protection Act 1998
- Freedom of Information Act 2000
- Human Rights Act 1998
- Statistics and Registration Services Act 2007
- Environmental Information Regulations 2004.

4.1.3.4 Quality Assurance

Quality control of data is an integral part of all research and takes place at various stages, during data collection, data entry or digitization, and data checking. It is vital to develop suitable procedures before data gathering starts. During data collection, researchers must ensure that the data recorded are accurate. The quality of data collection methods used strongly influences data quality, and documenting in detail how data are collected provides evidence of such quality. Digital collection is recommended to minimize human error in data collection.

Quality control measures during data collection can include:

- calibration of instruments to check the precision, bias and/or scale of measurement
- taking multiple measurements, observations or samples
- checking the truth of the record with an expert
- using standardized methods and protocols for capturing observations, alongside recording forms with clear instructions immediately after collection
- computer-assisted interview software to: standardize interviews, verify response consistency, route and customize questions so that only appropriate questions are asked, confirm responses against previous answers where appropriate and detect inadmissible responses.

When data are digitized, transcribed, entered in a database or spreadsheet, or coded, quality is ensured by standardized and consistent procedures for data entry with clear instructions. This may include:

- setting up validation rules or input masks in data entry software using data entry screens
- using controlled vocabularies, anthologies, code lists and choice lists to minimize human generated variation
- detailed labelling of variable and record names to avoid confusion.

Data quality can be assured though data editing, cleaning, verification, cross-checking and validation. Such checking typically involves both automated and manual procedures:

- double-checking coding of observations or responses and out-of-range values
- checking data completeness
- adding variable and value labels where appropriate
- verifying random samples of the digital data against the original data
- double entry of data
- statistical analyses such as frequencies, means, ranges or clustering to detect errors and anomalous values

- correcting errors made during transcription
- peer review.

4.1.3.5 Data Backup

It is the responsibility of a researcher to ensure that their research data is regularly backed-up and stored securely for the life of the project and throughout the retention period.
Most Centers already have a back-up policy for data held on institutional storage and shared network spaces usually wholly managed by the ICT unit. Where possible, it would therefore be more efficient to have the research data backup process managed by these ICT units governed under service level agreements with Research units and Data management teams. Offsite backup storage is also encouraged to ensure that research data is sufficiently distant from the primary data center in the event the primary site is destroyed. These offsite facilities may be physical or cloud based.

4.1.3.6 Data Security

Physical security, network security and security of computer systems and files all need to be considered to ensure security of data and prevent unauthorized access, changes to data, disclosure or destruction of data. Data security arrangements need to be proportionate to the nature of the data and the risks involved. Attention to security is also needed when data are to be destroyed.
Data transmission and encryption: Transmitting data between locations or within research teams can be challenging for data management infrastructure. To ensure sensitive or personal data is secure for transmitting it must be encrypted to an appropriate standard. Only data confirmed as anonymized or non-sensitive should be transmitted in unencrypted form. Encryption maintains the security of data during transmission.
Data disposal: Having a strategy for reliably erasing data files is a critical component of managing data securely and is relevant at various stages in the data cycle. During research, copies of data files no longer needed can be destroyed. At the conclusion of research, data files which are not to be preserved need to be disposed of securely.

4.1.3.7 Data Sharing

Research data are a valuable resource, usually requiring much time and money to be produced. Many data have a significant value beyond usage for the original research as is being increasingly recognized by donors. Some reasons for sharing data include:

- encourages scientific enquiry and debate
- promotes innovation and potential new data uses
- leads to new collaborations between data users and data creators
- maximizes transparency and accountability
- enables scrutiny of research findings
- encourages the improvement and validation of research methods

- reduces the cost of duplicating data collection
- increases the impact and visibility of research
- promotes the research that created the data and its outcomes
- can provide a direct credit to the researcher as a research output in its own right
- provides important resources for education and training.

As Centers become keen to share research data to increase the impact and visibility of their research, below, there are three 3 main areas that Centers should consider:

- Access control
- Data citations
- Ethics, consent and legal considerations.

D*raft tools to operationalize pre-open access data sharing* exist and are shared by the System Management Office of CGIAR System Organization (SMO) and can be accessed through the link https://gender.cgiar.org/webinar-data-sharing

4.1.3.8 Access Control

Different levels of access can be placed on research data. Access levels range from entirely open through to restricted access. It is, therefore, important to consider where and how the data will be managed for the longer term as there need to be systems in place to protect confidentiality and manage access.

A) **Open data**

When research data is open, it is freely available to use, reuse and redistribute. Potential re-users of research data need to have clear guidance about what they can and cannot do with the data. This is normally achieved via a license.

Article 4.1.5 of the CGIAR OADM policy states that "Suitable open licenses shall be used that recognize the legal rights to information products and encourage their use and adaptation."

The license conditions upon which information products are made Open Access may vary depending on the nature of the information products and the need to limit or restrict access or usage rights to certain audiences and users. No single license is appropriate for all research projects. A link to the CGIAR guide to choosing an appropriate license can be found here https://docs.google.com/viewer?a=v&pid=sites&srcid=Y2d4Y2hhbmdlLm9yZ3xvYWQtc3VwcG9ydC1wYWNrfGd4OjcxNjkzYTQ5MDlkZTYzZGY .

Creative Commons licenses (CC) provide free, easy-to-use copyright and machine-readable licenses to make a simple and standardized way to define permission to share and use creative work taking into account specific data ownership needs or restrictions. The most commonly recommended of these by donors is entirely unrestricted for re-use (CC-BY). There are other licenses available for software etc. e.g. GNU.

B) **Restricted Data**

When research data is restricted, it is not freely available to use, reuse and redistribute. Access restrictions can be applied on data at the repository level through use of password-controlled access to individual resource. However, it is still possible to allow for discoverability and global awareness of the research by making the metadata openly available.

According to the CGIAR OADM implementation guidelines: The general principle is to make information products Open Access, but that is always "subject to the legal rights and legitimate interest of stakeholders and third parties, including intellectual property rights, confidentiality, sensitivity (including price and politically-sensitive information), farmers' rights and privacy."

Exceptions are referred to in Articles 6.2, 6.3, and 6.4 of the CGIAR IA Principles.

Other ways of imposing access regulations for data may include:

- placing data under embargo for a given period of time until confidentiality is no longer pertinent
- providing secure access to data through enabling remote analysis of confidential data but excluding the ability to download data.

4.1.3.8 Data Citations

There is an increasing expectation that the outputs of publicly funded research, including the data should be made available for others to use. That means published data should be well-described (metadata), citable, discoverable and re-usable wherever possible.

As more journals have begun to require that data is made available to support research claims, research data has become a valuable asset that requires citation. The process of registration and citation depends on the type of data published and how it is published. Data can be considered as 'published' when it is generally discoverable.

Benefits of data citation include:

- Enables the sharing and re-use of datasets, which in turn has been shown to increase the author citation
- Makes data legitimately citable
- Acknowledges data ownership
- Allows for the use of data citation metrics.

The DOI system supports the citation of research data in scholarly communications and research data collaborations. A digital object identifier (DOI) is a unique alphanumeric string assigned by a registration agency (the International DOI Foundation) to identify content and provide a persistent link to its location on the Internet. The publisher assigns a DOI when an article is published and made available electronically. Once a project has finished a project management unit or equivalent will need to check whether the data generated by the project is accessible.

4.2 Institutional level activities

Whilst data is mainly generated through project activities, at an organizational level systems and processes should exist to support effective data management.

4.2.1 Governance and risk management

RDM cannot be effective in an organization without support of its governance structures. Such support may include:

- Championing effective RDM
- Approval of RDM design and related policies
- Allocation and approval of budgets for RDM activities
- Risk management processes assessing RDM related risks
- Communication to staff at all levels e.g. through regular updates or town-hall meetings, of the importance of effective RDM
- Setting up and monitoring performance indicators around RDM implementation and maintenance
- Support for regular audits/other reviews of RDM.

A) **Risk management**

Risk management is integral part of managing an organization or a project. It helps to avoid disasters and aids effective decision-making. Organizational/project activities are continuously assessed for exposure to risk including any risks related to data management.

Risk management is a three-step process that calls for:

A) Identification of the risks related to research data
B) Analyzing the risk in terms of likelihood of occurrence and the consequences
C) Putting in place mitigating actions to manage the risks.

For more information about the risk management, please refer to the Good Practice Notes on Risk Management.

B) **IT infrastructure**

IT is about effectively applying technology to the organizations data/information assets in order to help the organization reach its goals. IT includes the hardware, software and other facilities which underpin data-related activities. It plays a significant role in supporting data lifecycle and an organization in its efforts to manage one of its most valuable assets. Areas where IT department's support is vital include:

- Staff training on IT security measures
- IT network controls to support effective sharing of data
- Acquisition and maintenance of software to process, retain, share and archive data and data depositories
- Data access security controls
- Data back up
- Business continuity to ensure accessibility of data.

C) **Support services**

People and other means of providing advice and support, such as online toolkits.

4.2.3 Data Architecture

Functions, departments and units at a Center collect, maintain and utilize data for various purposes. To maximize the benefits of data, data sets should be well integrated through data architecture. Data architecture is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations.

The design and creation of modern data architectures is a process that brings in the whole enterprise, stimulating new ways of thinking, collaborating, and planning for data and information requirements. It's an opportunity for business decision makers to sit down with IT colleagues and figure out what kind of business they want to be in, what kinds of information they seek to propel that business forward, and what needs to be done to capture and harness that information. Architecture is more important than ever because it provides a road map for the organization to follow. Without a well-planned, careful, deliberate approach to data architecture, another type of architecture rises to take its place—a "spaghetti architecture" approach that occurs when every business unit or department sets out to buy its own solutions.

The essential components needed to build a modern data architecture are[3]:

- Identify the types of data that are most valuable and has high business impact. This data may have been within the organization's data environments for some time, but the means and technologies to surface such data, and draw insights, have been prohibitively expensive. Today's open source and cloud offerings enable Centers to pull and work with such data in a cost-effective way.
- Make data governance a priority. The process of identifying, ingesting, and building models for data needs to assure quality and relevance for the business. Responsibility for data must be established, whether it's individual data owners, committees, or centers of excellence.
- Design the data architecture to last. Data architecture should not be wedded to a particular technology or solution. If a new solution comes on the market, the architecture should be able to accommodate it. The types of data coming into enterprises can change, as do the tools and platforms that are put into place to handle them. The key is to design a data environment that can accommodate such change.
- Support real time access. A modern data architecture needs to be built to support the movement and analysis of data to decision-makers and at the right time it is needed. Also, it's important to focus on real-time from two perspectives. There is the need to facilitate real-time access to data, which could be historical; and there is the requirement to support data from events as they are occurring. For the first category, existing infrastructure such as data warehouses have a critical role to play. For the second, new approaches such as streaming analytics are critical. Data may be coming from transactional applications, as well as devices and sensors across the Internet of Things and mobile devices. A modern data architecture needs to support data movement at all speed whether it's sub-second speeds, or with 24-hour latency.

[3] Adapted from the article "8 Steps to Building a Modern Data Architecture" of the Data trends and applications magazine

- Consider security threats. A modern data architecture recognizes that threats are constantly emerging to data security, both externally and internally. Data managers and architects are in the best and most knowledgeable position to understand what is required for data security in today's environments.
- Develop a master data management (MDM) strategy. With a master data management repository, Centers have a single "gold copy" that synchronizes data to applications accessing that data. The need for an MDM-based architecture is critical, organizations are consistently going through changes. Often, organizations end up with data systems running in parallel, and often, critical records and information may be duplicated and overlap across these silos.
- Position data as a service. Many Centers have a range of databases and legacy environments, making it challenging to pull information from various sources. Access can be enabled through a virtualized data services layer that standardizes all data sources. Data as a service is by definition a form of internal cloud, in that data (along with accompanying data management platforms, tools, and application) are made available to the organization as reusable, standardized services. The potential advantage of data as a service is that processes and assets can be prepackaged based on corporate or compliance standards and made readily available within the organization's cloud.
- Offer self-service environment. With self-service, business users can configure their own queries and get the information or analyses they want, or conduct their own data discovery, without having to wait for their IT or data management departments to deliver the information. In the process, data application can reach and serve a larger audience than previous generations of more limited data applications. The route to self-service is providing front-end interfaces that are simply laid out and easy to use for business owners. In the process, a logical service layer can be developed that can be re-used across various projects, departments, and business units. IT still has an important role to play in a self-service-enabled architecture providing for security, monitoring, and data governance.

4.2.4 Training, awareness and continued support

Effective implementation of an institutional research data management framework requires that all institutional research staff receive adequate training on the Center's policies, procedures and practices with regards to research data management.

Induction training programs need to incorporate adequate information on RDM to enable all newly onboard research staff to be at par with existing staff on the Center's requirements towards proper research data management. Additionally, for longer projects, refresher training and inductions should be considered.

Training materials and self-read materials may be developed and made readily available for all research staff for quick references. However, the Center needs to ensure that the trainings are always kept up to date with the changes in research as well as institutional policies.

4.2.5 Continuous Monitoring and Feedback

A) **Policy compliance monitoring and RDM maturity**

Acknowledging the policies and guidelines as defined in the institution's research data management framework at the start of a research project is of little use in itself. Demonstrating compliance through review or audit frameworks allows non-compliance to be identified early and corrective actions to be taken. Compliance monitoring is a key component of any effective research data management process.

The Centers should include monitoring and auditing systems that are designed to detect variations in expected outcome or intentional deviations when an employee purposely seeks to stray from a defined process. Additionally, effective lines of communication with employees regarding compliance concerns, questions, or complaints need to be established.

B) **Impact and usage monitoring**

For research centers, the biggest output to the wider community are the research outputs and information products that are shared with the world. It is however useful for the Center to be able to track and measure the impact of their outputs. Over the past decade, institutions have been developing and testing ways in which they can best measure this impacts. One of the great benefits of publishing research data with proper identifiers is the ability to track usage and citation statistics.

According to the CGIAR Open Access and Data Management implementation plan 2014, Centers are encouraged to experiment with new ways of measuring, assessing, and tracking research outputs. At the minimum, Center data management plans should address: Which metrics are being collected, how these metrics are collected, and how they are interpreted in order to understand usage, impact, and uptake of materials disseminated through Open Access.

Impact is, in its figurative sense, the effect or influence that one agent, event or resource has on another. It is distinct from, but related to, concepts such as attention (how many people know about the entity) and dissemination (how widely a resource has been distributed). When considering proposed metrics, it is therefore important to consider what exactly is being measured and the strength of the evidence it provides for the impact of the entity in which one is interested:
No one metric can hope to represent fairly all possibilities, so it is worth exploring the variety of metrics that can be used. There are risks and concerns about reading too much into any given statistic, but metrics do provide an accessible way of uncovering evidence that might be suitable for use in an impact case study. Suggested metrics that can be used include Citations of data:

- Page views
- Downloads
- Social media links
- Post-publication peer review.

There are many impact measurement services that a Center may consider including:

- Thomson Reuters Data Citation Index

- ImpactStory
- PLoS Article-Level Metrics
- PlumX
- Altmetric
- ResearchGate Score
- Google Scholar and Microsoft Academic Search
- SocialCite
- PaperCritic
- ReaderMeter
- Crowdometer.

4.2.6 Funding agency considerations

Even as the CGIAR Centers continue to implement Research Data Management strategies and frameworks, it is important to note that funders have already begun realizing the vital importance of proper management and open sharing of data as they fund the creation of scholarly research, education and training materials and rich data with the public goods and far-reaching innovation and impact in mind.

Increasing number of funding agencies and donors have now adopted open access policies for their grant funded research which are then cascaded to funding agreements. These policies generally require the unrestricted access and re-use of all peer-reviewed published research, funded in whole or in part including any underlying data sets. Therefore, for the Centers to be able to keep up such is critical.
Below are links to some of these funders' policies.

	Funder Name	**Link to Policy**	**Effective from**
1	DFID	https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/181176/DFIDResearch-Open-and-Enhanced-Access-Policy.pdf	01 Nov 2012
2	USAID	http://www.globalcommunities.org/USAID-open-data-policy	01 Oct 2014
3	Bill & Melinda Gates Foundation	http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy	01 Jan 2015

5. BIBLIOGAPHY AND CREDITS

This GPN was developed under the leadership of Pierre Pradal, CGIAR IAU Director, by Beryl Akullah, CGIAR IAU Senior IT Auditor and Madina Bazarova, Associate Director, IAU with kind contributions from:

- Medha Devare, Senior Research Fellow, Program Support
- Peter Gardiner, Senior Manager, Program Support

It was based on the following materials:

- The DGI Data Governance DGI Data Governance Framework Framework, Prepared by Gwen Thomas, The Data Governance Institute published at http://www.datagovernance.com/wp-content/uploads/2014/11/dgi_framework.pdf

- Research data management in practice prepared by Mercury Solutions

- Carly Strasser. Research data management a primer publication of the National Information Standards Organization

- CGIAR (2013). CGIAR Open Access and Data Management Policy (the "Policy")

- CGIAR (2014). CGIAR Open Access and Data Management Implementation Guidelines

- CGIAR (2012). CGIAR Principles on the Management of Intellectual Assets ("CGIAR IA Principles")

- Data Sharing published at https://gender.cgiar.org/webinar-data-sharing/

- DAMA-DMBOK2 Framework by DAMA international published at https://www.dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf

- John Schmidt, information governance vs Data Governance published at http://blogs.informatica.com/2014/09/03/information-governance-vs-data-governance-who-cares/#fbid=IkOr5iue4l-

- Data management and Curation published at: https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/framework.html

- Max Wilkinson, Howard Amos, Lise Morton, Brian Flaherty, Shari Hearne, Helen Lynch, Heather Lamond, Natalie Dewson, Mike Kmiec, Janette Nicolle, Erin-Talia Skinner and Gillian Elliot; Research data management framework report published at http://www.universitiesnz.ac.nz/files/CONZUL-RDM%20Framework%20Report%202015%20FINAL.pdf

- Data management resources published at https://www.dataone.org/

- Guidelines for Best Practices in Data Management – Roles and Responsibilities published at https://www.for.gov.bc.ca/his/datadmin/DataMgmtRolesResp-2010Sept-FINAL-Approved.pdf

- Various data management resources published at http://www.data-archive.ac.uk/

- Various data management resources published at http://www.dcc.ac.uk/

- Various data management resources published at http://www.ands.org.au/guides/creating-a-data-management-framework

- Various data management resources published at http://guides.is.uwa.edu.au/c.php?g=325196&p=2177437

- Various data management resources published at www2.usgs.gov/datamanagement/preserve/repositories.php

- Creative Commons published at https://en.wikipedia.org/wiki/Creative_Commons_license

APPENDIX 1: EXISTING STANDARDS AND APPROACHES TO DATA MANAGEMENT

2.1 Acknowledged well known standards, methodologies and tools

Although we have not come across globally recognized standards specifically on research data management, for the purposes of this GPN, it would be important to look at 'The DAMA Data Management Body of Knowledge (DAMA-DMBOK)' framework. This framework provides some valuable insights on data management that can be effectively utilized in the design of robust research data management frameworks.

2.1.1 The DAMA Data Management Body of Knowledge (DAMA-DMBOK) framework

The Data Management Association (DAMA) is a not-for-profit, vendor-independent, international association of technical and business professionals dedicated to advancing the concepts and practices of information resource management (IRM) and data resource management (DRM). DAMA's primary purpose is to promote the understanding, development and practice of managing information and data as a key enterprise asset. The group is organized as a set of more than 40 chapters and members at large around the world, with an International Conference held every year.

The DAMA Guide to the Data Management Body of Knowledge" (DAMA-DMBOK Guide) was first published in April 5th, 2009 and there have been several revisions since then.

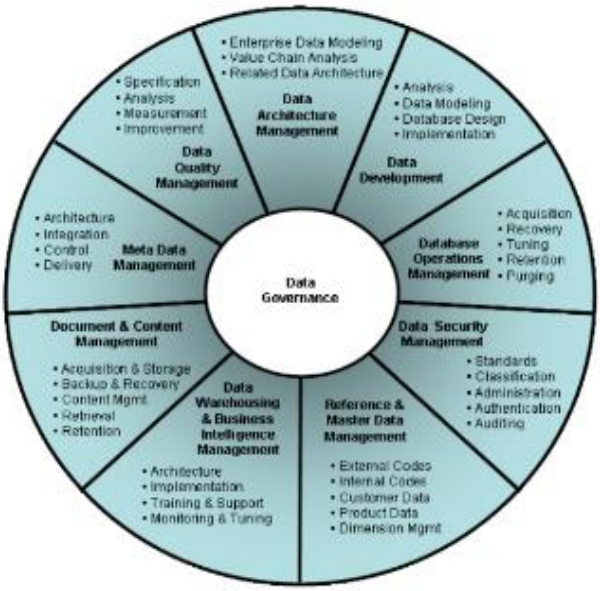
The DAMA-DMBOK Functional Framework Version 3 identifies 10 major Data Management Functions, each described through 7 Environmental Elements. The matrix below graphically presents DAMA-DMBOK Functional Framework and is a useful way of picturing the entire framework.

	Environmental Elements						
Data Management Functions	Goals & Principles	Activities	Deliverables	Roles & Responsibilities	Technology	Practices & Techniques	Organization & Culture
Data Governance							
Data Architecture Management							
Data Development							
Database Operations Management							
Data Security Management							
Reference & Master Data Management							
Data Warehousing & Business Intelligence Management							
Document & Content Management							
Meta Data Management							
Data Quality Management							

The DAMA-DMBOK Functional Framework, Version 3

2.1.1.1 Data Management Functions

The framework defines ten knowledge domains which are at the core of Information and Data Management.



Data management functions and the scope summary:

A) ***Data Governance***
The exercise of authority, control and shared decision-making (planning, monitoring and enforcement) over the management of data assets. Data Governance is high-level planning and control over data management.

B) ***Data Architecture Management***
The development and maintenance of enterprise data architecture, within the context of all enterprise architecture, and its connection with the application system solutions and projects that implement enterprise architecture.

C) **Data Development**
The data-focused activities within the system development lifecycle (SDLC), including data modeling and data requirements analysis, design, implementation and maintenance of databases data-related solution components.

D) **Database Operations Management**
Planning, control and support for structured data assets across the data lifecycle, from creation and acquisition through archival and purge.

E) **Data Security Management**
Planning, implementation and control activities to ensure privacy and confidentiality and to prevent unauthorized and inappropriate data access, creation or change.

F) **Reference & Master Data Management**
Planning, implementation and control activities to ensure consistency of contextual data values with a "golden version" of these data values.

G) **Data Warehousing & Business Intelligence Management**
Planning, implementation and control processes to provide decision support data and support knowledge workers engaged in reporting, query and analysis.

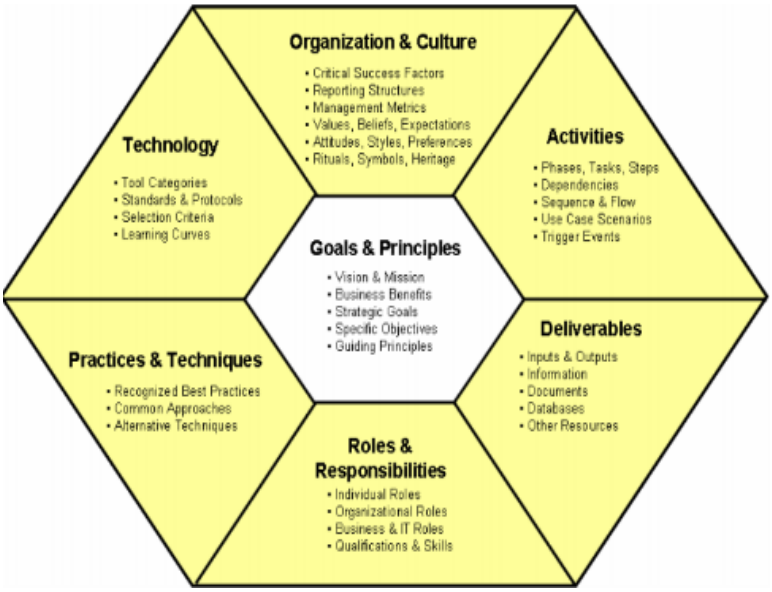
H) **Document & Content Management**
Planning, implementation and control activities to store, protect and access data found within electronic files and physical records (including text, graphics, image, audio, video)

I) **Meta Data Management**
Planning, implementation and control activities to enable easy access to high quality, integrated metadata.

J) **Data Quality Management**
Planning, implementation and control activities that apply quality management techniques to measure, assess, improve and ensure the fitness of data for use.

2.1.1.2 Environmental Elements

The 7 Environmental Elements provide a logical and consistent way to describe each function. The idea of environmental elements is not a new one. A commonly referenced structure identifies three elements: Process, Technology and People. However, for this framework there is deeper coverage of 2 elements i.e. *Process* (Processes, Deliverables, Principles, Methods & Techniques) and *People* (Roles & Responsibilities, Organizational & Cultural Issues).



The basic Environmental Elements are:

A) **Goals & Principles**: The directional business goals of each function and the fundamental principles that guide performance of each function.
B) **Activities**: Each function is further decomposed into lower level activities. Some activities are grouped into sub-functions. Activities can be further decomposed into tasks and steps.
C) **Deliverables**: The information and physical databases and documents created as interim and final outputs of each function. Some are considered essential, some are generally recommended, and others are optional depending on circumstances.
D) **Roles and Responsibilities**: The business and IT roles involved in performing and supervising the function and the specific responsibilities of each role in that function. Many roles will participate in multiple functions.

The supporting Environmental Elements are:

A) **Practices & Procedures**: Common and popular methods and techniques used to perform the processes and produce the deliverables. May also include common conventions, best practice recommendations and alternative approaches without elaboration.
B) **Technology:** Categories of supporting technology (primarily software tools), standards and protocols, product selection criteria and common learning curves. In accordance with DAMA policies, specific vendors or products should not be mentioned.
C) **Organization and Culture**: These issues might include:
    - Management Metrics – measures of size, effort, time, cost, quality, effectiveness, productivity, success and business value
    - Critical Success Factors
    - Reporting Structures
    - Contracting Strategies
    - Budgeting and Related Resource Allocation Issues
    - Teamwork and Group Dynamics
    - Authority & Empowerment
    - Shared Values & Beliefs
    - Expectations & Attitudes
    - Personal Style & Preference Differences
    - Cultural Rites, Rituals and Symbols
    - Organizational Heritage
    - Change Management Recommendations.

2.1.1.3 DAMA-DMBOK Functional Outline

For the 10 functions, there are 102 activities and there could be more than 12 activities within a function. Each activity is categorized as belonging to one of four Activity Groups:

A) ***Planning Activities*** **(P)** Activities that set the strategic and tactical course for other data management activities. Planning activities may be performed on a recurring basis.
B) ***Control Activities*** **(C)** Supervisory activities performed on an on-going basis.
C) ***Development Activities*** **(D)** Activities undertaken within projects and recognized as part of the systems development lifecycle (SDLC), creating data deliverables through analysis, design, building, testing and deployment.
D) ***Operational Activities (*****O)** Service and support activities performed on an on-going basis.

DAMA-DMBOK FUNCTION ACTIVITY TABLE

Function	Activities
1. Data Governance	1.1. Data Management Planning 1.1.1. Identify Strategic Enterprise Data Needs (P) 1.1.2. Develop & Maintain the Data Strategy (P) 1.1.3. Establish the Data Management Professional Organizations (P) 1.1.4. Identify & Appoint Data Stewards (P) 1.1.5. Establish Data Governance & Stewardship Organizations (P) 1.1.6. Develop, Review & Approve Data Policies, Standards and Procedures (P) 1.1.7. Review & Approve Data Architecture (P) 1.1.8. Plan and Sponsor Data Management Projects & Services (P) 1.1.9. Estimate Data Asset Value & Associated Data Management Costs (P) 1.2. Data Management Supervision & Control 1.2.1. Supervise the Data Management Professional Staff & Organizations (C) 1.2.2. Coordinate Data Governance Activities (C) 1.2.3. Manage & Resolve Data Related Issues (C) 1.2.4. Monitor & Ensure Regulatory Compliance (C) 1.2.5. Monitor Conformance with Data Policies, Standards and Architecture (C) 1.2.6. Oversee Data Management Projects & Services (C) 1.2.7. Communicate & Promote the Value of Data Assets (C)
2. Data Architecture Management	2.1. Develop & Maintain the Enterprise Data Model (P) 2.2. Analyze & Align with Other Business Models (P) 2.3. Define & Maintain the Data Technology Architecture (P) (same as 4.2.2) 2.4. Define & Maintain the Data Integration Architecture (P) (same as 6.2) 2.5. Define & Maintain the DW / BI Architecture (P) (same as 7.1.2) 2.6. Define & Maintain Enterprise Taxonomies (P) (same as 8.2) 2.7. Define & Maintain the Meta Data Architecture (P) (same as 9.2)
3. Data Development	3.1. Data Modeling, Analysis & Design 3.1.1. Analyze Information Requirements (D) 3.1.2. Develop & Maintain Conceptual Data Models (D) 3.1.3. Develop & Maintain Logical Data Models (D) 3.1.4. Develop & Maintain Physical Data Models (D)

	3.2. Detailed Data Design 3.2.1. Design Physical Databases (D) 3.2.2. Design Related Data Structures (D) 3.2.3. Design Information Products (D) 3.2.4. Design Data Access Services (D) 3.3. Data Model & Design Quality Management 3.3.1. Develop Data Modeling & Database Design Standards (P) 3.3.2. Review Data Model & Database Design Quality (C) 3.3.3. Manage Data Model Versioning and Integration (C) 3.4. Data Implementation 3.4.1. Create & Maintain Development & Test Databases (D) 3.4.2. Create & Maintain Test Data (D) 3.4.3. Migrate & Convert Data 3.4.4. Build & Test Information Products ((D) 3.4.5. Build & Test Data Access Services (D) 3.4.6. Build & Test Data Integration Services (D) 3.4.7. Validate Information Requirements (D) 3.4.8. Prepare for Data Deployment (D)
4. Database Operations Management	4.1. Database Support 4.1.1. Implement & Maintain Database Environments (C) 4.1.2. Implement & Control Database Changes (C) 4.1.3. Acquire Externally Sourced Data (O) 4.1.4. Plan for Data Recovery (P) 4.1.5. Backup & Recover Data (O) 4.1.6. Set Database Performance Service Levels (P) 4.1.7. Monitor & Tune Database Performance (O) 4.1.8. Plan for Data Retention (P) 4.1.9. Archive, Retrieve and Purge Data (O) 4.1.10. Manage Specialized Databases (O) 4.2. Data Technology Management 4.2.1. Understand Data Technology Requirements (P) 4.2.2. Define the Data Technology Architecture (P) (same as 2.3) 4.2.3. Evaluate Data Technology (P) 4.2.4. Install & Administer Data Technology (O) 4.2.4. Inventory & Track Data Technology Licenses (C) 4.2.5. Support Data Technology Usage & Issues (O)
5. Data Security Management	5.1. Understand Data Privacy, Confidentiality & Security Needs (P) 5.2. Define Data Privacy & Confidentiality Policies & Standards (P) 5.3. Define Password Standards & Procedures (P) 5.4. Design & Implement Data Security Controls (D) 5.5. Manage Users, Passwords & Group Membership (C) 5.6. Manage Data Access Views (C)

	5.7. Manage Data Access Permissions (C) 5.8. Monitor User Authentication & Access Behavior (C) 5.9. Classify Information Confidentiality (C) 5.10. Audit Data Security (C)
6. Reference & Master Data Management	6.1. Understand Reference & Master Data Integration Needs (P) 6.2. Define the Data Integration Architecture (P) (same as 2.4) 6.3. Implement Reference & Master Data Management Solutions (D) 6.4. Control Code Values & Other Reference Data (C) 6.5. Integrate Master Data (O) 6.6. Replicate Reference and Master Data (O) 6.7. Maintain Dimensional Hierarchies (O)
7. Data Warehousing & Business Intelligence Management	7.1. Data Warehousing & Business Intelligence Planning 7.1.1. Understand Business Intelligence Data Needs (P) 7.1.2. Define & Maintain the DW / BI Architecture (P) (same as 2.5) 7.2. Data Warehousing & Business Intelligence Implementation 7.2.1. Implement Data Warehouses & Data Marts (D) 7.2.2. Implement Business Intelligence Tools & User Interfaces (D) 7.2.3. Implement Enterprise Reporting (D) 7.2.4. Implement Management Dashboards & Scorecards (D) 7.2.5. Implement Analytic Applications (D) 7.3. Data Warehousing & Business Intelligence Support 7.4. Train Business Professionals (O) 7.5. Replicate & Transform Data for Business Intelligence (O) 7.6. Monitor & Tune Data Warehousing Processes (C) 7.7. Support Business Intelligence Activity (O) 7.8. Monitor & Tune BI Activity and Performance (C)
8. Document & Content Management	8.1. Plan for Managing Electronic & Physical Documents (P) 8.2. Define & Maintain Enterprise Taxonomies (P) (same as 2.6) 8.3. Implement & Maintain Document Storage Systems (D) 8.4. Acquire & Store Documents (O) 8.5. Index Document Information Contents (O) 8.6. Backup & Recover Documents (O) 8.6. Support Document Content Analysis (O) 8.7. Support Document Access, Circulation & Update (O) 8.8. Monitor & Tune Document Retrieval Performance (C) 8.9. Archive, Retrieve & Purge Documents (O) 8.10. Audit Document & Content Management (C)
9. Meta Data Management	9.1. Understand Meta Data Requirements (P) 9.2. Define the Meta Data Architecture (P) (same as 2.7) 9.3. Develop & Maintain Meta Data Standards (P) 9.4. Implement a Managed Meta Data Environment (D) 9.5. Create & Maintain Meta Data (O) 9.6. Integrate Meta Data (C) 9.7. Manage Meta Data Repositories (C)

	9.8. Distribute & Deliver Meta Data (C) 9.9. Support Meta Data Reporting and Analysis (O)
10. Data Quality Management	10.1. Develop and Promote Data Quality Awareness (O) 10.2. Profile, Analyze & Assess Data Quality (D) 10.3. Define Data Quality Requirements & Business Rules (D) 10.4. Test & Validate Data Quality Requirements (D) 10.5. Define Data Quality Metrics & Service Levels (P) 10.6. Measure & Monitor Data Quality (C) 10.7. Manage Data Quality Issues (C) 10.8. Correct Data Quality Defects (O) 10.9. Design & Implement Operational DQM Procedures (D) 10.10. Monitor Operational DQM Procedures & Performance (C) 10.11. Audit Data Quality (C)

2.1.2 The Data Governance Institute Data Governance framework

As governance is a key element in any management strategy, for the purpose of this GPN, we shall also highlight the DGI Data governance framework developed by the Data Governance Institute.

The Data Governance Institute (DGI) is the industry's oldest and best known source of in-depth, vendor-neutral Data Governance best practices and guidance. Founded in 2003 by Gwen Thomas, primary author of the DGI Data Governance Framework, it's the number one recognized name in the industry, with practitioners around the world consistently reporting that they have based their programs on the DGI Data Governance Framework and supporting materials. DGI introduced the DGI Data Governance Framework in 2004 in response to an emerging need for a way to classify, organize, and communicate complex activities involved in making decisions about and taking action on enterprise data.
What is data governance? Data Governance is the exercise of decision-making and authority for data-related matters. Put in another way, it is a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods.

The difference between data governance and IT governance is the following:
IT governance is the primary way that stakeholders can ensure that investments in IT create business value and contribute toward meeting business objectives. It consists of following established frameworks e.g. **CobiT®, ITIL, ValIT®,** and **ISO 38500** and best practices to gain the most leverage and benefit out of IT investments and support accomplishment of business objectives.
Data governance consists of the processes, methods, tools, and techniques to ensure that data is of high quality, reliable, and unique (not duplicated), re-usable so that downstream uses are more trusted and accurate. Hence it seeks to ensure efforts are put in formal management controls to govern critical data assets. Data governance accounts for all aspects of data both unstructured and structured.
In essence, data governance should be a part of the IT governance program.

Data Governance requires input from Subject Matter Experts and management representatives that understand data. Although both data and information may be classified as corporate assets, management of each and the focus areas are different. IT Governance groups focus may be on the IT Portfolio Management issues e.g. performance of systems, installing of new applications etc. and not data specific issues such as the ones below:

A) ***Data transparency*** - what data do we have in the enterprise, where is it, and how is it secured.

B) ***Data lineage*** - what is the system of record for various types of data, how does it move between systems, and what transformations were applied in the process.

C) ***Data Quality*** - what rules can be applied systematically in the capture, monitoring, and measurement of data assets.

D) ***Service Levels*** - what are the required service levels for the timeliness of data delivery or synchronization between copies of the data.

E) ***Data Security*** - how can data be kept secure regardless whether it is controlled by an application system, copied to a test or training database, or stored in the cloud.

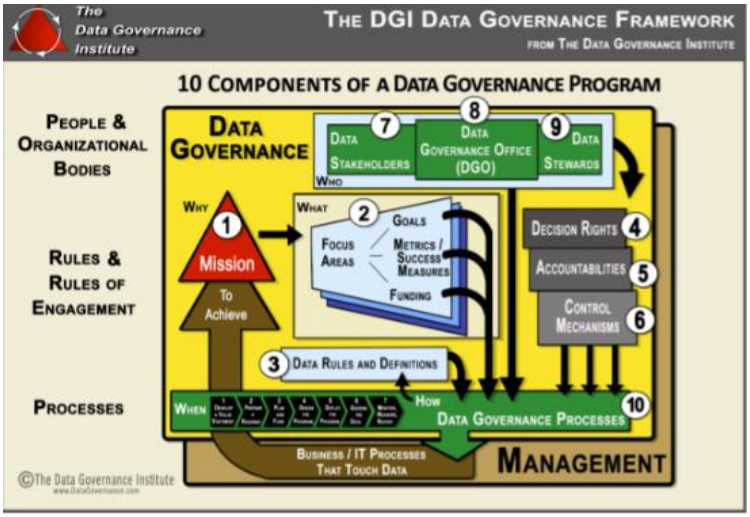
F) ***Change Impact*** - what is the impact on existing and historical data and data processes if a given system change is implemented.

G) ***Data Ownership*** - who is accountable for maintaining and operating the data stores, whether they are stand-alone copies or linked to a production application.

It should be noted that in most centers data management teams are separate from IT teams. It is therefore vital to ensure that roles and responsibilities by each team are well defined due to the overlapping nature of activities. This avoids conflicts while at the same time encourages efficiency in the use of skills and resources.

The aim of the DGI data governance framework is to standardize data definitions across an enterprise or initiative.

According to the framework, there are 10 universal components of a Data Governance program as shown below:



The DGI Framework

2.1.2.1 Data Governance Components that deal with 'Rules and Rules of Engagement'

Component #1: Mission and Vision

At its highest level, Data Governance typically has a three-part mission:

- Proactively define/align rules.
- Provide on-going, boundary-spanning protection and services to data stakeholders.
- React to and resolve issues arising from non-compliance with rules.

Along with the mission, there is need for a clear vision. What could your organization look like with a mature Data Governance program? How about without one?

Component #2: Goals, Governance Metrics / Success Measures, Funding Strategies

Just like goals should be SMART - Specific, Measurable, Actionable, Relevant, and Timely. Everyone involved in Data Governance should know what success looks like, and how it's being measured. Clear value statements help one consider funding options available for the program.

Component #3: Data Rules and Definitions

This component refers to data-related policies, standards, compliance requirements, business rules, and data definitions. Depending on the organizations focus areas, the program may work to:

- Create new rules/definitions
- Gather existing rules/definitions
- Address gaps and overlaps
- Align and prioritize conflicting rules/definitions
- Establish or formalize rules for when certain definitions apply.

Component #4: Decision Rights

Before any rule is created or any data-related decision is made, a prior decision must be addressed: who gets to make the decision, and when, and using what process? It is the responsibility of the Data Governance program to facilitate the collection of decision rights that are the "metadata" of data-related decisions.

Component #5: Accountabilities

Once a rule is created or a data-related decision is made, the organization will be ready to act on it. Who should do what, and when? For activities that do not neatly map to departmental responsibilities, the Data Governance program may be expected to define accountabilities that can be baked into everyday processes.

Component #6: Controls

It's well established that data is constantly at risk. With the proliferation of sensitive data breaches – and the consequences for those who were entrusted with the data, it is becoming clear that data can also represent risk.

Often the Data Governance program is asked to recommend data-related controls that could be applied at multiple levels of the controls stack (network / operating system; database; application; user processes) to support governance goals.

2.1.2.2 Data Governance Components that Deal with People and Organizational Bodies

Component #7: Data Stakeholders

Data Stakeholders come from across the organization. They include groups who create data, those who use data and those who set rules and requirements for data. Because Data Stakeholders affect and are affected by data-related decisions, they will have expectations that must be addressed by the Data Governance program. Some will expect to be included in some kinds of data-related decisions. Some will be expected to be consulted before decisions are formalized, and others will be satisfied to be informed of decisions after they are made.

Component #8: A Data Governance Office (DGO)

The Data Governance Office (DGO) facilitates and supports these governance activities. It collects metrics and success measures and reports on them to data stakeholders. It provides ongoing Stakeholder Care in the form of communication, access to information, record-keeping and education/support.

Component #9: Data Stewards

The Data Stewardship group consists of a set of Data Stakeholders who come together to make data-related decisions. They may set policy and specify standards, or they may craft recommendations that are acted on by a higher-level Data Governance Board.

Data Governance programs with a focus on Data Quality may also include Data Quality Stewards. These roles typically report to a business function or Data Quality team, with dotted-line accountabilities to Data Governance. These stewards examine sets of data against criteria for completeness, correctness, and integrity. They make corrections as appropriate and refer other issues to the DGO.

2.1.2.3 The Process of Governing Data

Component #10: Proactive, Reactive, and Ongoing Data Governance Processes

Components 1-6 of the DGI Data Governance Framework deal with rules. They also describe the "rules of engagement" employed by components 7-9 (People and Organizational Bodies) during governance. This last component, Processes, describes the methods used to govern data. Ideally, these processes should be standardized, documented, and repeatable. They should be crafted in such a way to support regulatory and compliance requirements for Data Management, Privacy, Security, and Access Management.

Every organization will decide how much structure and formality to bring to the process of governing data. The Data Governance Institute recommends (and routinely implements) formal, documented, repeatable procedures for:

- Aligning policies, requirements, and controls
- Establishing decision rights
- Establishing accountability

- Performing stewardship
- Managing change
- Defining data
- Resolving issues
- Specifying data quality requirements
- Building governance into technology
- Stakeholder care
- Communications
- Measuring and reporting value.

2.2 Recognized practices used by other similar organizations

Research organizations and Universities are the main organization whose core activities result in the production of valuable research data across different spectra of research disciplines. While each may have specific approach to how management of their data is undertaken the general concepts apply across all fields. We have included a Research Data Management Framework report authored by the CONZUL working group published at:
http ://www.universitiesnz.ac.nz/files/CONZULRDM%20Framework%20Report%202015%20FINAL.pdf.
CONZUL stands for “The Council of New Zealand University Librarians” and is a Committee of 8 Universities in New Zealand. CONZUL's objective is to act collectively to improve access for students and staff of New Zealand universities to the information resources required to advance teaching, learning and research.

APPENDIX 2: ACRONYMS

CC – Creative commons

COBIT – Control Objectives for Information Technology

CONZUL - The Council of New Zealand University Librarians

CRP - CGIAR Research Programs

DAMA - Data Management Association

DFID – Department of international development

DGI – Data Governance Institute

DGO – Data Governance Organization

DMBOK - Data Management Body of Knowledge

DMP – Data Management Plan

IA – Intellectual Assets

ICT – Information, Communication and Technology

ISO – International Standards Organization

OADM – Open Access and Data Management

RDM – Research Data Management

SMO – System Management Office

SRF - Strategy and Results Framework

USAID – United States Agency of International Development