ASSOCIAÇÃO BRASILEIRA DE ENERGIA NUCLEAR – ABEN
2019 International Nuclear Atlantic Conference - INAC 2019
Santos, SP, Brazil, October 21-25, 2019

DATA GOVERNANCE FOR NEUTRON ACTIVATION ANALYSIS

Claudia Chamas, Edson Gonçalves Moreira and Mário Olímpio de Menezes.

Nuclear and Energy Research Institute (IPEN - CNEN/SP)
Research Reactor Center (CERPq)
Av. Prof. Lineu Prestes 2242
05508-000 São Paulo, SP, Brazil
cchamas@usp.br

ABSTRACT

This study describes the importance of managing scientific research data, and incorporating the demands that are currently presented such as data sharing, data repositories, metadata management, as well as ensuring the truth and value of data. A proper data management helps in gaining insight on data governance. The area of study is the Nuclear and Energy Research Institute (IPEN – CNEN/SP) where there is a wide variety of researches, yielding lots of data. IPEN has already recognized the importance of proper management of research data as an essential part of the research's good practices and of the institution's commitment to the truthfulness and quality of its scientific research. Following this commitment, IPEN has established the necessity of data management and sharing in order to ensure the greatest possible benefits for scientific and technological advancement. A case study with data from neutron activation analysis experiments, belonging to the Neutron Activation Analysis Laboratory (LAN-CERPq) of IPEN is presented. The current status of data management was assessed and a basic implementation plan for a proper data management was proposed, delineating also a path to the adoption of data governance practices at a larger scale in the whole IPEN-CNEN/SP.

Keywords: data management, scientific research, data governance, activation analysis

1. INTRODUCTION

Data Management Plans are recognized as an important component of the proper management of research data and have been demanded for more than a decade by the leading government or private research funding agencies around the world (North America, Europe, Australia, etc).

Following the trending in the worldwide scientific practice, FAPESP[1], (The São Paulo Research Foundation), started to require from researchers applying for funding in some

1 Available at: http://www.fapesp.br/en/about

categories or specific calls, the presentation of a Data Management Plan to take care of all data generated during the project and also stating how those data will be preserved and shared in the long term after the end of the project.
Therefore, it is considered necessary that the data resulting from projects funded by FAPESP be managed and shared in a way that guarantees the greatest possible benefit for scientific and technological progress, besides rationalizing resources [1].

Appropriate management and sharing of such data facilitates research's reproducibility while also promotes new researches, thanks to the possibility of reuse of those data, as well as help in carrying out new analyzes. Sharing of scientific data also facilitates the training of new generations of researchers and the exploration of topics not foreseen in the original project.

The Data Management Plan is a text that should answer two basic questions:

1. Which data will be generated by the project?
2. How will they be preserved and made available, considering ethical, legal, confidentiality and other issues?

Some FAPESP calls can specify the desired format of the Plan. For all other cases, the Plan submitted as an annex to a proposal to FAPESP must be a text of up to two pages containing the following information.

a) Description of the data and metadata produced by the project - for example, samples, collection records, forms, models, experimental results, software, graphics, maps, videos, spreadsheets, audio recordings, databases, teaching materials and others.

b) When applicable, legal or ethical restrictions for sharing such data, policies to ensure privacy, confidentiality, security, intellectual property and others.

c) Preservation policy and sharing (for example, immediate sharing or only after acceptance of the associated publication). Grace period (before sharing) and period during which data will be preserved and made available.

d) Description of mechanisms, formats and standards to store such items in a way that makes them accessible by third parties. This description may include the use of repositories and services from other institutions.

Several other templates for Data Management Plans can be found on the Internet. The instructions vary greatly, depending on the area of knowledge. However, the set of basic information to be provided is always the same - what data will be produced by the project, sharing constraints, how they will be shared, and how they will be preserved.

A data management plan (DMP) should basically describe how the data will be processed during a project and what happens to the data after the project ends, thus first identifying what data to be processed. [2]

These plans typically cover all or part of the data life cycle - from data discovery, collection and organization (e.g. spreadsheets, databases), through quality assurance /

quality control, documentation (e.g. data, laboratory methods) and use of data to preserve and share data with others.

There are 10 simple rules that help ensure that the data is secure and shareable [2]:
1. Determine the research sponsor requirements
2. Identify the data to be collected
3. Define how the data will be organized
4. Explain how data will be documented
5. Describe how data quality will be assured
6. Presenting a Strategy for Storage and Preservation of Sound Data
7. Define project data policies
8. Describe how the data will be released
9. Assign Roles and Responsibilities
10. Prepare a realistic budget

As an example, the National Science Foundation (NSF)[3] , an independent federal agency created by the US Congress in 1950, requires both, a data management and a research product sharing plans, including preservation, documentation and data sharing, samples, physical collections, curriculum and other related research and education products that should be described in the Special Information and Supplemental Documentation section of the proposal.

The same for National Institutes of Health[4], USA which presents a Practical Guide to Creating a Data Sharing Plan, showing a detailed description of how to prepare a plan, its content, and the importance of data sharing for the advancement of science. Although specific to Health area, the text is applicable to a wide range of areas of knowledge. 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 reuse. [3]

Also as a document to be of interest to the planning of data management is the appointment of the repository to store research data. The data should be submitted to repositories recognized by the community, specific discipline, where possible, or general repositories, if there are adequate resources available in the community.

Important to consider is the process of digital curation to deal with the planning, evaluation and reassessment of the digital object, which encompasses digital preservation as part of its cycle; curation is a key aspect of a data management plan. [4]

Digital curation is the active involvement of information professionals in management, including the preservation of digital data for future use. While there have been people doing different aspects of data curation and digital preservation for decades, recent events have brought various ideas, organizations and individuals to focus more intensely on digital curation. Reports in the US by the National Science Foundation and the

2 Available at (portuguese only): http://www.fapesp.br/gestaodedados/
3 Chapter II - Instructions for Preparation of Proposals (NSF)
4 National Institutes of Health, EUA - https://grants.nih.gov/grants/oer.htm

American Council of Learned Societies and in the UK by Dr. Liz Lyon of UKOLN[5][6]have pointed out the aspects of digital curation that need to be in place to ensure that digital objects can be maintained, preserved and remain available for future use.

Increasingly, "digital curation is becoming a comprehensive concept that includes digital preservation, data curation, electronic records management, and digital asset management" [5]

In short, the management of scientific data requires detailed plans showing compliance with the institution's policies, security concerns, the curation and the indication of the repositories of such data. [6]

Data management clarifies and determines the purposes of the data. For data to be presented accurately, consistently, and securely, data governance is required.

1.1 Data Management and Governance

According to the Data Management Body of Knowledge (DMBOK), data management is the planning, execution and supervision of policies, practices and projects that acquire, control, protect, deliver and improve the value of data and assets [7].

This means the encapsulation of data management and data governance within business functions. The DMBOK is the most complete work in data management, constituting therefore, as a base reference for this research. Its purpose is to effectively teach how to control and leverage the use of data assets and its mission and objectives are meet and exceed the information needs of all stakeholders in terms of availability, safety and quality of data.

Data management is a responsibility of both the Information Technology sector of a company and all its internal and external customers. It goes from top management, where data is used in the generation of strategic information, to operational level professionals, who are often responsible for collecting and producing data [8].

Data management, whether for a company, research or teaching institution, is the work that aims, among other facets, to unify the people, processes and systems that manage and protect the data, through a consistent way to exchange data, with understanding and processing to increase quality and trust.

Data governance has a number of definitions across the literature; it accounts for planning, supervising and controlling the management and use of data; it is the exercise of authority, control and decision making (DMBOK).

5 Dr Liz Lyon is the Director of UKOLN at the University of Bath UK, where she leads work to promote synergies between digital libraries and open science environments. Available at: http://www.dcc.ac.uk/about-us/dcc-staff-directory/liz-lyon

6 Available at: http://www.dcc.ac.uk/digital-curation/what-digital-curation

The National Association of State Chief Information Officers (NASCIO)[7] argues that data governance is essential to ensure accuracy, sharing, and data protection. Data governance is "the organizational framework, rules, right decisions and responsibilities of people and information systems regarding the performance of information-related processes" [9].

For the Data Governance Institute, data governance "is a system of decision-making rights and responsibilities for information-related processes, executed according to agreed models. These processes describe who can take what actions, with what information, and when, under what circumstances, using what methods" [10]

The main understanding is that data governance should consider data protection, data liability, processes or flows through which data are processed, and specific procedures; It should also consider statements of how to perform these procedures, and who will ensure that governance happens. In addition, it should help to standardize the policy and technology that understands the data as assets of the organization.

Another important issue related to data governance, encompasses the decision-making framework, that is, the "specification of the model for decision-making rights and responsibilities to encourage the desirable behavior in the use of data" [11].

The Nuclear and Energy Research Institute (IPEN - CNEN/SP) recognizes the importance of proper management of research data as an essential part of good research practice. As part of this vision, IPEN instituted a working group that aims to guide its researchers on the best data management practices required by the funding agencies.

This act corroborates the academic works that are already being developed under the scope of the e-Science project fostered by IPEN. The present work is being developed under the guidance of e-Science project, focusing at the Neutron Activation Analysis Laboratory, specifically working with data from neutron activation experiments. It's important to mention that other analytical techniques are employed at the Lab, such as Atomic Absorption Spectrometry, but the data generated by such techniques are out of scope of the present work. The data management model being developed here should be easily replicated in other laboratories at IPEN or other research centers.

Data management activities can be viewed as a cycle, encompassing the research cycle. Starting with the Research Question and the Research Planning, reusable data is searched, a data management plan is then prepared, followed by the data collection, analysis, archiving, publication and again going to new research question. This cycle is shown on Figure 1.

7 Founded in 1969, NASCIO is a non-profit association that represents executives of information andinformation technology. The top members of the state are senior state government officials who have executive and state responsibility for leadership in information technology. Available at: https://www.nascio.org/

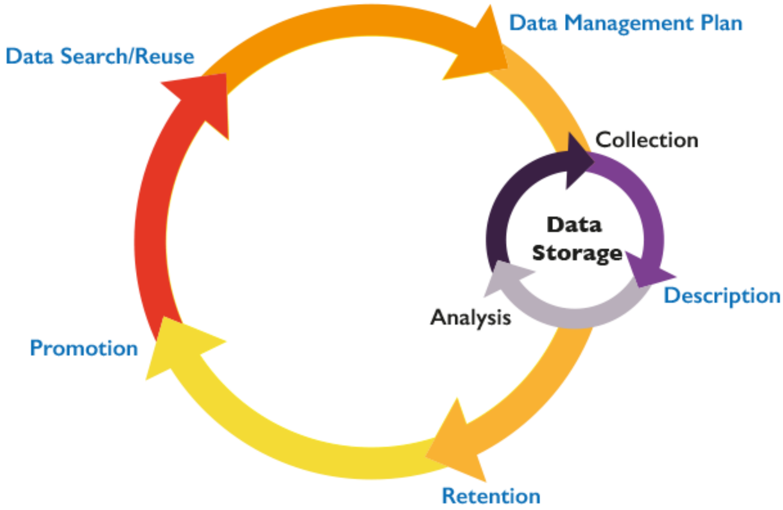

Figure 1: Research Data Management Lifecycle, from The University of California, Santa Cruz, Data Management LibGuide[8]

1.2. Neutron Activation Analysis (NAA)

Neutron Activation Analysis (NAA) consists of the induction of radioactivity in a sample and subsequent detection and quantification of the induced radioactivity. The induction of radioactivity is usually carried out by the activation of the sample with neutrons in a nuclear reactor and, through reactions of neutron capture, a fraction of the atomic nuclei present in the sample is transformed into radioactive isotopes, that decay in time according to their half-lives, producing the emission of gamma rays with characteristic energies.

The detection of these gamma rays via spectrometry allows the identification and quantification of the various activated elements present in the sample. The application of the technique is done basically through two approaches: a comparative and a parametric one. In the first, standards and samples are irradiated simultaneously and the measurements of the radioactivities induced in the samples are compared with those of the standards allowing the quantification of the elements. In the second, the contents of the analyzed elements are found through calculations involving physical parameters of the
reactor used and the target elements [12].

NAA technique can be further categorized as:

(a) Prompt gamma rays analysis, where the measurement occurs during irradiation. This technique is known as PGAA;

8 Available at: http://guides.library.ucsc.edu/datamanagement

(b) Analysis by activation with neutrons of gamma rays of decay ("delayed gamma rays"), which is the technique used at LAN-CERPq (IPEN).

The present study presents first steps in the establishment of data governance for the Neutron Activation Analysis Laboratory (LAN – CERPq), whose main analytical technique is the Instrumental Neutron Activation Analysis (INAA); this study aims to serve as a guiding instrument for scientific data governance best practices.

2. MATERIALS AND METHODS

As said above, data management is a required practice for scientific researchers. A established data management culture makes easier the implementation of data governance.

1. Identify the data generated by neutron activation analysis experiments;
2. Describe precisely these data, uniformizing the naming used by researchers;
3. Identify the main data, required to identify the quality of the analysis;
4. Identify the key data required to demonstrate the quality of the analysis.

Thus, careful readings were carried out in dissertations, thesis, articles and operational procedures (OP) produced by researchers and students on neutron activation experiments in order to identify the various data produced by the technique.

A detailed analysis of each entry of the resulting collection was conducted, aiming at eliminating duplicates (coming from different nomenclatures for the same data), as well as organizing these data according to the typical steps of Instrumental Neutron Activation Analysis (INAA).

A listing of these data (in a spreadsheet) together with their description and typical associated values was presented to the researchers of the LAN-CERPq to be validated and completed, if any data was missing.

After this phase, the complete INAA data listing, organized according to its typical steps, was obtained and are presented in Table 1.

One of the most demanded aspect of scientific data is its quality. The higher the quality of data, the greater the impact that its utilization can achieve. Higher data quality is increasingly demanded to foster the advancement of science.

Ensuring data quality is one of the essential aspects of a data management plan (DMP). As such, the INAA data compilation was once again presented to the experts, asking them to answer two questions: "Which data identifies the quality of the INAA?" and "Which is the most important data in the INAA, that is, data that can be designated as key or essential data?"

The obtained answers were tabulated and the results are presented in Table 2.

Table 1 - INAA data validated by LAN-CERPq experts

	Data Description	Question 1	Question 2
		This data identify the quality of the INAA	**Which is the most important data in the INAA, that is, the key data?**
	PROJECT (Dissertation, Thesis, etc)		
1	Title	x (example)	x (example)
2	Subject / Analytical problem to be addressed		
3	Author		
4	Advisor		
5	Institute		
6	Laboratory		
7	Start date		
8	Closing date		
9	Collaborator		
10	Great area of study		
11	Subarea		
	PROCESS		
12	Sample (ex: geological)		
13	Representative sample		
14	Sample description		
15	Collection date		
16	Location of collection / area		
17	Seasons		
18	Points of study (north, south, etc.)		
19	Stocking and storaging		
20	Sample storage temperature		
21	Preparation of the sample (drying, milling, etc.)		
22	Percentage of sample moisture		
23	Stocking / storing the prepared sample		
	CERTIFIED REFERENCE MATERIAL DATA		
24	Name of the reference material		
25	Certificate of reference material		
26	Certification report for ref. mat.		
27	Percentage of moisture in the CRM		
	PREPARATION OF THE		

	SYNTHETIC STANDARD		
28	Standard solution certificate value		
29	Diluted volume		
30	Final volume		
31	Volume pipetted		
32	Mass of elements		
	IRRADIATION		
33	Sample mass		
34	Type of beam (thermal, epidermal)		
35	Irradiation casing		
36	Irradiation time		
37	Irradiation casing (casings, capsules)		
38	Irradiation card number		
	RADIATION MEASUREMENTS		
39	Decay time		
40	Measurement date / hour		
41	Minimum sample and standard counting time		
42	Measuring equipment data		
43	Software / program name		
44	Gamma spectrum file		
	QUANTITATIVE / QUALITATIVE SAMPLE ANALYSIS		
45	Radionuclide data used in the calculation		
46	Blank analysis data		
47	Calculation file: peak listing and counting data for matrix		
48	Detection limits results		
	RESULTS		
49	CRM concentration results (individual, mean, relative error)		
50	Synthetic Standard concentration results (individual, mean, relative error)		
51	Sample concentration results (individual, mean, relative error)		

Table 2: Data that can be used to assess the research quality in INAA

Data Description	Num of votes for Question 1	Num of votes for Question 2
PROJECT (Dissertation, Thesis, etc)		
Title		
Subject / Analytical problem to be addressed	3	
Worker	3	
Advisor if any	1	
Institute	1	1
Laboratory	3	
Start Date	2	
Closing Date	2	
Collaborator if any	2	
Great area of study	1	
Subarea	1	1
PROCESS		
Sample (ex: geological)		1
Representative sample	2	2
Sample description	2	3
collection date	3	1
Location of collection / area	3	2
Seasons	4	
Points of study (north, south, etc.)	3	
Storage	3	2
Sample storage temperature	4	2
Preparation of the sample (drying, milling, etc.)	4	3
Percentage of sample moisture	4	4
Stocking / storing the prepared sample	4	1
CERTIFIED REFERENCE MATERIAL DATA		
Name of the reference material	4	4
Certificate of reference material	4	3
Certification report for ref. mat.	3	2
Percentage of moisture in the CRM	3	1
PREPARATION OF THE SYNTHETIC STANDARD		
Standard solution certificate value	1	3
Diluted volume	4	2
Final volume	3	3
Volume pipetted	3	2
Mass of elements	3	3

IRRADIATION		
Sample mass	1	3
Type of beam (thermal, epidermal)	4	1
Irradiation casing (casings, capsules)	3	2
Irradiation time	3	
Irradiation card number	3	2
RADIATION MEASUREMENTS		
Decay time		
Measurement date / hour	3	4
Measuring equipment data	4	1
Software / program name	4	2
Gamma spectrum file	4	2
QUANTITATIVE / QUALITATIVE SAMPLE ANALYSIS		
Radionuclide data used in the calculation		3
Blank analysis data	3	3
Calculation file: peak listing and counting data for matrix	4	4
Detection limits results		
RESULTS		
CRM results (individual, mean, relative error)	3	4
Synthetic Standard concentration results (individual, mean, relative error)	3	4
Sample concentration results (individual, mean, relative error)	3	4

3. RESULTS AND DISCUSSION

The consolidation of all data pertaining to the INAA in the LAN-CERPq/IPEN, resulted in the identification of 51 (fifty one) feature data items, presented in Table 1. In addition to being the first step for the establishment of data management, it forms a good basis of INAA knowledge, identifying its main processes. Knowing and understanding the data, its order and the clarification of its origin, leads to an better understanding of the information and relations behind the data.

The analysis of the answers to the two questions, whose tabulation is shown in Table 2, allows the establishment of a hierarchical structure for the data quality framework. Using an arbitrary threshold of 2 votes, 46 (forty six) data entries that positively indicates the quality of a INAA experiment were selected. For the second question, using the same threshold, 27 (twenty seven) data entries were selected, indicating those considered key data for a INAA experiment.

An additional result of this analysis is the identification of a minimum data set of INAA experiments that should never be overlooked concerning its quality; this can be used as guidance to researchers, specially students, to help them in the process of getting better data and consequently, better results of their experiments.

Other key issues to ensure data quality are those related to data exchange between researchers when working on distributed or multisite networks. Concerns with file naming conventions and, eventually, versioning applications, should be taken as mandatories, together with a well established metadata management. [12]

The next stages of this study will expand on some of the 10 simple rules presented above for the construction of a data management plan, addressing practical issues required for the establishment of a data management culture. Some aspects are related to technological solutions (software and hardware) and other are simply habit changes or new habit adoption.

An important aspect to any organization aiming to implement data governance is to achieve people engagement in data management plans construction.

4. CONCLUSIONS

A good data governance framework helps to ensure that data be formally identified and described and also supports the assurance of quality standards, as well as ensuring accuracy, sharing and data protection, among other goals.

Data quality aspects influence the reliability, completeness, accuracy and consistency of all data produced by a research project; it's also the main basis of data management plans. As evidenced by this study, the use of a controlled vocabulary corroborates the quality of the data, providing a uniform naming across researchers and facilitating data sharing. This is an important achievement, since one of the measures of data quality is its usability [14]. The more usable the data the bigger is quality.

Our study has identified a subset of INAA data that can be used to assess the quality of the research; this was done by applying a specific questionnaire to the specialists; the resulting feature set is shown on Table 2. By mapping each entry of this table to the corresponding step in the neutron activation analysis, it's possible to implement specific actions to improve data acquisition quality in each step, having as a result the improvement of the overall research quality.

ACKNOWLEDGMENTS

To NAA specialists participating in this research and LAN-CERPq / IPEN

REFERENCES

1. FAPESP, "Code of Good Scientific Practices", available at www.fapesp.br (2014)

2. W.K. Michener, “Ten Simple Rules for Creating a Good Data Management Plan”, *PLOS Computational Biology*, **vol. 11,** n. 10, e1004525, out (2015)
3. M.D. Wilkinson, et.al, “The FAIR Guiding Principles for scientific data management and stedwarship”, *Scientific Data*, Nature Publishing Group, **vol 3**, n. 160018 (2016)
4. S.A. Siebra, V.R. Borba, M. Galindo, M.K.F.O. Miranda, L.L.L. Tavares, J.N.N. Oliveira, “Curadoria digital: além da questão da preservação digital”. In: *Encontro Nacional De Pesquisa Em Ciência Da Informação*, Florianópolis. UFSC, 2013. Available at: http://www.egov.ufsc.br/portal/sites/default/files/curadoria_digital_0.pdf
5. E. Yakel, “Digital curation”, OCLC Systems & Services, **vol. 23** , n o 4, p. 335–340, nov (2007)
6. G. W. Ryan and H. R. Bernard, “Data Management and Analysis Methods”, *Handbook of Qualitative Research*, 2nd Ed., Norman Densin and Yvonna Lincoln, Eds. Thousand Oaks, CA: Sage Publications, 2000, p. 769-802
7. M. Mosley, M. H. Brackett, “The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK Guide)”, 1st Edition. Bradley Beach, N.J.: Technics Publications, 2010.
8. “Definitions of Data Governance”, *The Data Governance Institute* , Homepage: http://www.datagovernance.com/adg_data_governance_definition/ (2019)
9. G. Thomas, “Alpha Males and Data Disasters The Case for Data Governance”. Orlando, FL, USA, Brass Cannon Press (2006)
10. D.I.T. Fávaro, C. Afonso, M.B.A. Vasconcellos, S.M.F. Cozzolino, “Determinação de elementos minerais e traços por ativação neutrônica, em refeições servidas no restaurante da Faculdade de Saúde Pública/USP”. *Ciênc. Tecnol. Aliment.*, **v. 20**, n. 2, p. 176-182, Aug. 2000 .
11. K. Weber and B. Otto, “A Contingency Approach to Data Governance”, *ACM Journal of Data and Information Quality*, **Vol. 1**, No. 1, Article 4, June 2009
12. S. Jones, “How to Develop a Data Management and Sharing Plan”, *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides (2011)
13. M. G. M. C. Semmler, “Biomonitoring of Hg, Cd, Pb and other elements in coastal areas of the State of São Paulo through the transplanted mussel - Perna perna (Linnaeus, 1758)”, *Journal of Radioanalytical and Nuclear Chemistry,* **Vol. 278,** pp 547-551, (2008)
14. L.L. Pipino, Y.W. Lee, R.Y. Wang. “Data Quality Assessment”, *Communication of the ACM*. **Vol. 45**, No. 4 (2002)