A framework for data governance and management of the Semear Digital Center

Debora Pignatari Drucker[43], Cássia Isabel Costa Mendes[44], Isaque Vacari[45], Daniel Rodrigo de Freitas Apolinário[46], Luciana Alvim Santos Romani[47]

Abstract

Data is the basis of scientific research and is increasingly important to support decision-making, becoming even more critical in a center focused on research and development of digital technologies for agriculture. The Center of Science for Development in Digital Agriculture (Semear Digital Center) generates heterogeneous data in terms of structure, volume, frequency, and content, as well as its comprehensive temporal, spatial, and thematic scopes. Our objective is to present a framework for data governance and management for the Semear Digital Center, developed to ensure data preservation, documentation, integrity, accountability, and regulatory compliance throughout its life cycle, from data generation to publication. We performed an exploratory study based on a literature review, document analysis, and expert consultation. Our framework for data governance carefully followed the legal regulations, as well as the sponsor´s and the leader institution's rules. We developed a data lifecycle model to guide the data management framework that will assist researchers and other stakeholders in properly manage data assets generated within the Center's scope.

Keywords: Data Management; Data Governance; Data Lifecycle; Agriculture.

1. Introduction

Data is the basis of research, and it is increasingly important that it can be read and understood by both humans and machines (Richard et al., 2023). Ensuring its integrity is also crucial for decision-making in many areas, including agricultural production. Therefore, good data management practices throughout its life cycle have become very

[43]Embrapa Digital Agriculture. ORCID: 0000-0003-4177-1322. Email: debora.drucker@embrapa.br.
[44]Embrapa Digital Agriculture. ORCID: 0000-0002-7646-5870 Email: cassia.mendes@embrapa.br.
[45]Embrapa Digital Agriculture. ORCID: 0000-0002-8719-1107 Email: isaque.vacari@embrapa.br.
[46]Embrapa Digital Agriculture. ORCID: 0000-0002-7636-536X. Email: daniel.apolinario@embrapa.br.
[47]Embrapa Digital Agriculture. ORCID:. 0000-0002-7386-3515. Email: luciana.romani@embrapa.br.

relevant in scientific research (Sinaeepourfard et al., 2016), becoming even more critical in a center focused on research and development of digital technologies. Additionally, developing digital agriculture technologies based on artificial intelligence depends on good-quality data to train the models (Barbedo et al., 2024).

Given the importance of applying good practices of data governance and management in scientific and technological endevours, our objective is to present a framework for data governance and management developed for ensuring data preservation, documentation, integrity, accountability and regulatory compliance in the context of the Center of Science for Development in Digital Agriculture (Semear Digital Center) (Semear Digital, 2025) throughout its life cycle, from data generation to publication.

2. Methods

To develop a data governance and management framework for the Semear Digital Center, we performed an exploratory study based on a literature review, document analysis, and expert consultation.

The Center of Science for Development in Digital Agriculture is an initiative led by Embrapa and encompasses more than 100 researchers and students from 7 institutions. It is organized into six axes of action: 1) Agrotechinological District (DAT) Selection and Impact Evaluation; 2) Connectivity; 3) AI and Remote Sensing; 4) Precision Agriculture and Automation; 5) Traceability; and 6) Partnerships and Communication. Since 2023, the Center has been generating heterogeneous data in terms of structure, volume, frequency, and content, as well as its comprehensive temporal, spatial, and thematic scopes.

The data governance framework was developed based on the study of the legal framework, i.e., the digital law applicable to data governance; the sponsor's data management policy; Embrapa's data policy, and its institutional repository data license. The data management framework was built nested in the data governance framework, considering the extensive data life cycle management literature. Several data life cycle models have been proposed to efficiently organize large and complex data sets, from creation to consumption, for effective data usage (Sinaeepourfard et al., 2016). We developed the framework considering the data life cycle from generation to publication and taking into account the particularities and challenges of the Semear Digital Center, including the diversity of research groups, institutions, and stakeholders involved. We

consulted with experts in cyberinfrastructure and cybersecurity to guarantee that our framework will ensure data preservation, documentation, integrity, accountability, and regulatory compliance.

3. Results and Discussion

3.1 Data Governance Framework

The governance practices of the Semear Digital Center include the dimension of legal regulation for data processing throughout its lifecycle. Data governance regulation is on the international and national legislative agenda, which is inserted in a broader globalized context related to the legal framework applied to the digital environment, and has repercussions for countries and public and private institutions. We can mention legislative initiatives from the European Union (EU) and Brazil in a comparative law dimension. The EU is cited as being at the forefront of drafting laws to regulate the digital environment and protect personal and non-personal data. In this sense, the General Data Protection Regulation (GDPR) – 2016/679 establishes rules for protecting personal data and its free circulation. It is essential to highlight that although the expression “data protection” is used, what is being protected is individuals, that is, data subjects. Not only are personal data subject to protection in the EU, but non-personal data is as well. In this sense, two regulations stand out, as shown in Table 1.

Following the trend of the EU, Mendes et al. (2023) explain that, with regard to data protection regulation, in Brazil the legal system on data has been consolidating, with highlights including the Federal Constitution, the Access to Information Law and the Internet Civil Rights Framework and the General Personal Data Protection Law – LGPD, which was inspired by the GDPR. The guidelines of the LGPD are applied within the scope of Semear Digital Center. In order to collect personal data, the data subject, usually the rural producer, is asked to sign a consent form authorizing the processing of their personal data and the agricultural data of their rural property. The purpose, adequacy and necessity, guidelines inherent to the LGPD, are also met, considering that, in Semear Digital Center, the purpose for collecting personal data is defined and communicated to the data subject, the processing of the data is adequate and not excessive concerning the purpose, and the collection is carried out considering only the data necessary to achieve the purpose. The guidelines align with the Center's objective, which is to overcome inequalities in the field through research, development, and innovation (RD&I) in

Information and Communication Technology (ICT), aiming to increase the production and productivity of small and medium-sized producers.

Table 1. European directives regulating data protection. Source: prepared by the authors.

Regulation	Description
Regulation (EU) 2018/1807 for the Free Flow of Non-Personal Data	- Ensures the free flow of data other than personal data - It sets out rules on data localisation requirements, data availability to competent authorities, and data portability for professional users
Data Governance Act (Regulation (EU) 2022/868)	- Establishes conditions for the reuse of data by public sector bodies - Defines a notification and supervision regime for the provision of data intermediation services

In addition to applying the regulation on the processing of personal data prescribed in the LGPD, data generated under the umbrella of the Semear Digital Center is obligated to comply with the sponsor´s Data Management Policy. The São Paulo Research Foundation, Fapesp, has a Data Management Policy (Fundação de Amparo à Pesquisa do Estado de São Paulo, 2025) that attests that “data resulting from projects financed by the Foundation be managed and shared in a way that guarantees the most significant possible benefit for scientific, technological, socioeconomic, and cultural advancement.” Fapesp requires a data management plan to be included in every grant proposal as a mandatory supplemental document, and its fulfillment is checked upon the grant´s reviews and reports. Besides, Embrapa Digital Agriculture, the host institution of the Center, has its Data Governance Policy (Embrapa, 2019), which aims “to strengthen the mechanisms for generating, organizing, processing, accessing, preserving, recovering, disseminating, sharing, and reusing Embrapa's informational assets”. The policy establishes guidelines for implementing processes to promote the FAIR principles (Wilkinson et al., 2016), and one of the means of its implementation is depositing data in the institutional research data repository, Redape[48]. Once data is published at Redape, a persistent identifier and a CC-BY-NC (Creative Commons, 2025) license are assigned to the dataset.

[48]Available at: https://www.redape.dados.embrapa.br/.

3.2 Data Management Framework

Considering the literature review, document analysis, and expert consultation, we defined a data lifecycle model for the Semear Digital Center, illustrated in Figure 1 and composed of the following phases and steps: Pre-processing phase: 1 - Data Planning, 2 - Data Collection and 3 - Raw Data Preservation and Documentation; Processing and Post-Processing phases: 4 - Data Assuring, Integration and Analysis, 5 - Data Description and 6 - Data Publication.

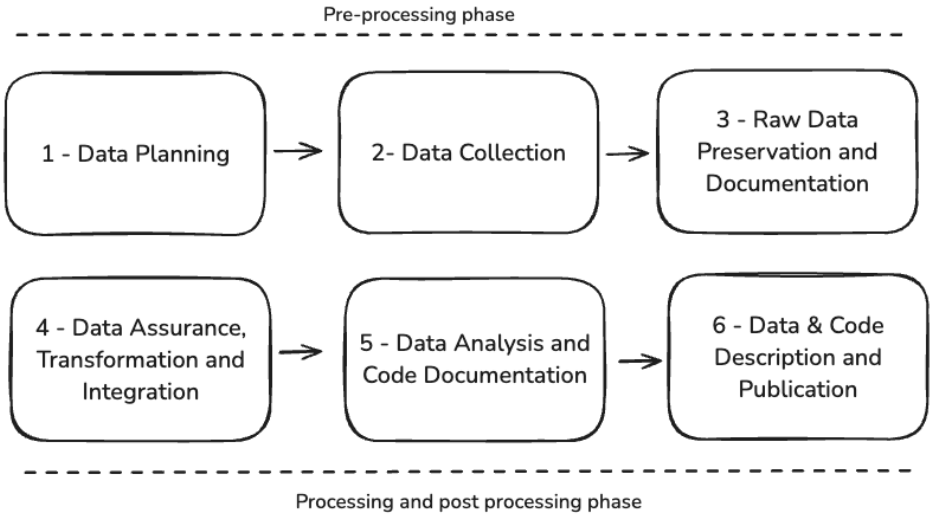

Figure 1. Data Life Cycle of the Semear Digital Center: the pre-processing phase is represented in steps 1 to 3, while the processing and post-processing phases are represented in steps 4 to 6.

The pre-processing phase is composed of the following steps. The Data Planning step was written during the proposal submission phase of the Semear Digital Center and is continuously developed by the different research groups within the Center, which also conduct the Data Collection step. In order to ensure data integrity and preservation, a cloud-based storage was structured in the Agrodigital infrastructure for the Raw Data Preservation and Documentation step. At this step, the data is accompanied by a basic documentation file, composed of the answers to the following: What was measured? Who measured it? When was it measured? Where was it measured? How was it measured? How is the data structured? Why was the data collected? Who should get

credit for this data (researcher AND funding agency)? How can this data be reused (licensing)?

The processing phase comprises the Data Assurance, Transformation, and Integration step, when the research groups perform a series of procedures to assure the data quality, validate, combine, and integrate the data. The following step in the processing phase is data analysis and code documentation. Once the final results are generated, which will be used in scientific publications and other research products, the post-processing phase is initiated, which is composed of the Data & Code Description and Publication step. To comply with the governance framework, the data and code, when applicable, will be published at Embrapa's institutional repository (Redape), accompanied by the metadata that describes them, based on the documentation file generated in the previous steps.

4. Conclusion

We presented the data governance and management framework developed to the Semear Digital Center to ensure data preservation, integrity, and accountability, in accordance with the applicable laws and regulations. Our approach took into account the heterogeneous data generated by the Center in terms of structure, volume, frequency, and content, as well as its temporal, spatial, and thematic scopes. Considering the complexity of the stakeholders involved in the digital agriculture solutions being developed, the legal framework related to data governance needed to be carefully followed, as well as the sponsor´s and the leader institution's regulations. The proposed data lifecycle model will assist researchers and other stakeholders in properly manage data assets generated within the Center's scope.

Future work includes developing case studies within the Center´s research groups, i.e., in the six axes of action, to detect possible limitations and make adjustments. Our approach contributes to increasing the impact, visibility, and credibility of the researcher, the research, and the institutions involved with the Semear Digital Center.

Acknowledgements

The authors thank Fapesp (Proc. 2022/09319-9) for the funding and the experts consulted.

References

BARBEDO, J. G. A.; ROMANI, L. A. S.; DRUCKER, D. P.; MACÁRIO, C. G. do N. Perspectivas dos agrodados e da digitalização da agricultura. In: MENDES, C. I. C.; MARANHÃO, J. de S. de A.; SARAIVA, A. M. (ed.). **Agricultura digital, agrodados e regulação**. Brasília, DF: Embrapa, 2024. cap. 5, p. 107–117. Available at: https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1170060/1/LV-Perspectivas-Agrodados-2024.pdf. Accessed on: March 10, 2025.

CREATIVE COMMONS. **Attribution-NonCommercial 4.0 International.** Mountain View. Available at: https://creativecommons.org/licenses/by-nc/4.0/. Accessed on: April 14, 2025

EMBRAPA. Resolução Consad nº 184, de 4 de abril de 2019. Política de Governança de Dados, Informação e Conhecimento da Embrapa. **Boletim de Comunicações Administrativas**, ano 45, n. 16, p. 1-19, 5 abr. 2019. Available at: https://www.embrapa.br/politica-de-governanca-de-dados-informacao-e-conhecimento. Accessed on: March 10, 2025

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO. **Gestão de dados**. Available at: https://fapesp.br/gestaodedados. Accessed on: March 10, 2025

MENDES, C. I. C.; BERTIN, P. R. B.; COSTA, M. M. Programa de governança em privacidade e proteção de dados pessoais na Administração Pública Federal. **Administração de Empresas em Revista**, v. 2, n. 32, e-6367, abr./jun. 2023. Available at: https://www.alice.cnptia.embrapa.br/alice/handle/doc/1154244. Accessed on: April 14, 2025

RICHARD, S.; GREGORY, A.; HODSON, S.; FILS, D.; KANJALA, C.; BELL, D.; WINSTANLEY, P.; EDWARDS, M.; HEUS, P.; BRICKLEY, D.; RIZZOLO, F.; MAXWELL, L.; LUIS, G.; BUTTIGIEG, P. L.; LE FRANC, Y. **Cross Domain Interoperability Framework (CDIF)**: discovery module (v01 draft for public consultation) (Versão 01). 2023. Zenodo. DOI: https://doi.org/10.5281/zenodo.10252564.

SEMEAR DIGITAL. **O Centro de Ciências para o Desenvolvimento em Agricultura Digital**. Available at: https://www.semear-digital.cnptia.embrapa.br/. Accessed on: March 10, 2025

SINAEEPOURFARD, A.; GARCIA, J.; MASIP-BRUIN, X.; MARÍN-TORDER, E. Towards a comprehensive data lifecycle model for big data environments. In: IEEE/ACM INTERNATIONAL CONFERENCE ON BIG DATA COMPUTING, APPLICATIONS AND TECHNOLOGIES (BDCAT '16), 3., 2016, Shanghai. **Proceedings** [...]. Los Alamitos: IEEE, 2016. p. 100-106. DOI: https://doi.org/10.1145/3006299.3006311.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J. W.; SANTOS, L. B. da S.; BOURNE, P. E.; BOUWMAN, J.; BROOKES, A. J.; CLARK, T.; CROSAS, M.; DILLO, I.; DUMON, O.; EDMUNDS, S.; EVELO, C. T.; FINKERS, R.; GONZALEZ-BELTRAN, A.; GRAY, A. J. G.; GROTH, P.; GOBLE, C.; GRETHE, J. S.; HERINGA, J.; ´t HOEN, P. A. C.; HOOFT,. R.; KUHN, T.; KOK, R.; KOK, J.; LUSHER, S. J.; MARTONE, M. E.; MONS, A.; PACKER, A. L.; PERSSON, B.; ROCCA-SERRA, P.; ROOS, M.; SCHAIK, R. van; SANSONE, S. A.; SCHULTES, E.; SENGSTAG, T.; SLATER, T.; STRAWN, G.; SWERTZ, M. A.; THOMPSON, M.; VAN DER LEI, J.; VAN MULLIGEN, E.; VELTEROP, J.; WAAGMEESTER, A.; WITTENBURG, P.; WOLSTENCROFT, K.; ZHAO, J.; MONS, B. The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, v. 13, p. 1-8, Mar. 2016. DOI: https://doi.org/10.1038/sdata.2016.18