NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Design and Implementation of a Data Repository for Collaborative Research: A Pilot Platform for Dataset Management and Continuity

Supporting Sustainable Research Through Structured Data Access and Collaboration

Ilyass Jannah

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Design and Implementation of a Data Repository for Collaborative Research: A Pilot Platform for Dataset Management and Continuity

Supporting Sustainable Research Through Structured Data Access and Collaboration

by

Ilyass Jannah

Master Thesis presented as a partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Flavio Luis Portas Pinheiro, MSc, NOVA Information Management School

July, 2025

Statement of Integrity

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, 2025]

Jannah Ilyass

ACKNOWLEDGEMENTS

I would like to express my appreciation with honesty to the respondents that have made a great contribution to this research project during its implementation.

First of all, my supervisor, Professor Flavio Luis Portas Pinheiro, has been constantly shaping the course of the current thesis through his mentorship, revisionary input, and academic advice. His proficiency in the field of knowledge has been invaluable.

I also owe thanks to my colleagues in the university whose intellectual interactions, intellectual challenge and methodological know-how have polished the theoretical model as well as the computing prototype. Their combined knowledge has helped me understand the practical issues of data-science.

I would like to recognize a special award to my close friend Achraf because he kept me enthusiastic and kept me going. His technical brainstorming, positive debate and constant support has been invaluable.

Last but not least, I would like to express my utmost gratitude to my family. Their encouragement, patience and faith provided the psychological and logistical solidity necessary to make this project a reality even when going through tough times. Their sacrifice made everything possible, me being here is thanks to them.

It was the combination of these people who gave structure, criticism, technical input, intellectual stimulation and emotional grounding to make this research possible.

Abstract

With researchers now putting so much weight on data driven methods, datasets are becoming fundamental assets that lead the way to innovation. Thus making them reproducible, collaborative and lasting is a main concern and storing, versioning, and organizing them should not be taken lightly. This project consists of the design and development of a centralized data repository application that is suited for academic and research environments, and gives solutions that are tailored for researcher students. The platform offers a well-designed, secure and easy-to-use interface to store various research datasets with a focus on version control, documentation of datasets and role-based access control. Built using modern web technologies such as NextJS for the backbone of the app, MongoDB for data storage, and mechanisms for versioning, the application bridges a major gap between data management practices and the needs of research teams. With making research processes more transparent and researchers more accountable in mind, this project introduces primary functions such as dataset metadata tracking, API-based retrieval of the data, and role controlled access. These tools essentially make the process of documentation easy while providing an extra layer of security. The thing that is actually interesting about this project is that it leads our field to reproducible research - which is a big deal in most contemporary scientific investigations. Combined with the standardization of documentation, the strength of version tracking, and the safety of the permission scheme, such a repository offers a good, long-term scaffolding to future collaborations. In a nutshell, the project provides students with the working tools they need to handle, share, and develop their data sets more efficiently, encouraging openness, trust, and sharing across each other.

Keywords

Data Management; Data Repository; Data Reproducibility; Centralized Storage; Data Sharing

Sustainable Development Goals (SDG):

Table of Contents

List of Figures

List of Tables

1. Introduction

In this era's IT landscape, data became an important cornerstone. This is all the more true for academic environments where being able to manage, curate, and share research data/results is crucial. Digital research output grows at an astonishing rate, putting more pressure on institutions forcing them to face challenges of data reliability while ensuring data preservation, promoting transparency, reproducibility, and collaboration. Research Data Management(RDM) became all the more important due to its key role in helping during the research lifecycle by enabling compliance with funding agency requirements and giving the researchers works more visibility and reuse. (Dunie, M., 2017)(Briney, K. A., Coates, H. L., & Goben, A., 2020)(Willoughby, C., & Frey, J. G., 2022)

But, it's not all sunshine and rainbows, the growth and availability of data comes with its challenges and struggles, particularly in the Global South where the implementation of a structured RDM framework still has its hurdles and the health sector where no real solution has been developed. Some of the most common obstacles are the scarcity of centralized policies, technical capacity, training, and standardized procedures for handling datasets(Mushi, G. E., Pienaar, H., & van Deventer, M., 2020). And it's not just the logistical aspect of it all, but it's more profound than that; it affects researchers' ability for secure data storage, responsible sharing, and documentation for future use. Fragmented practices like ad hoc naming conventions and unclear data ownership has been determined by multiple studies as one of the issues hindering the accessibility of data thus harming the research outputs (Kumar, A., Gawande, A., Paliwal, J., Pendse, V., Kale, S., Agarwal, A., ... & Raibagkar, S., 2025)(Briney, K. A., Coates, H. L., & Goben, A., 2020)(Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., ... & Burke, D. S., 2014).

These institutional barriers are worsened by the broader technological and strategic challenges associated with big data. In some domains like energy, health, and environmental sciences, researchers need more, larger, heterogeneous data sources making the need for scalable, high-performance data infrastructures more crucial than ever (Daki, H., El Hannani, A., Aqqal, A., Haidine, A., & Dahbi, A., 2017)(Almeida, F., & Calistru, C., 2013). For example, in smart grid systems, pairing the management of high velocity data streams with the promise of highly qualitative and interoperable data is essential to making the project succeed (Daki, H., El Hannani, A., Aqqal, A., Haidine, A., & Dahbi, A., 2017). But current technologies still are not up to that level yet, they still lack some key features like robust metadata, version control, and access protocols that would lead to consistent interpretation and reuse.

Once adding to the two previously mentioned dimensions, there is also the ethical side of things that make the problem more complex. Concerns make researchers more hesitant to share data causing stalls in innovation and hindrance in cross sectoral data collaboration which are all the more important for solving more intricate societal challenges (Metsis, S., 2020). Some of these concerns are intellectual property, misuse, and lack of recognition that

although literature emphasizes the importance of a transparent data ecosystem, still mainly depend on the human motive (Kumar, A., Gawande, A., Paliwal, J., Pendse, V., Kale, S., Agarwal, A., ... & Raibagkar, S., 2025)(Hulsen, T., 2020). Indeed, an effective collaboration requires the tools necessary to accomplish that such as interoperable platforms, clear role definitions, and confidence in data integrity that are still not available in research environments (Toma, M., Bönisch, C., Löhnhardt, B., Kelm, M., Bohnenberger, H., Winkelmann, S., ... & Kesztyüs, T., 2024).

Coming back to our university's context, most of the research students face similar problems while working on their thesis, mainly the need for data collection that wastes a lot of precious time and leads to data reuse as multiple students can have similar topics and end up with pretty much the same dataset as a starting point. Also as the university is pretty open and international, we can find students with different backgrounds and levels of expertise so the available tools might not be to the taste of everyone. (Wiehe, S. E., Rosenman, M. B., Chartash, D., Lipscomb, E. R., Nelson, T. L., Magee, L. A., ... & Aalsma, M. C., 2018)

Another side to look at is the institution(University) side. Although some advanced data management systems exist and can be used to solve some of the issues the students face, they reveal another set of problems.

- ➔ The use of such systems can become too costly for the university to handle.
- ➔ These systems lack the customizability and flexibility needed to be fully adapted and integrated into a student research environment.
- ➔ Most of them are not so easy to use and would need the student to learn more about them before harnessing their benefits making them sound way less appealing.
- ➔ They cause risks of vendor lock-in which is a situation where the customer becomes heavily reliant on a specific vendor for products and services, making it difficult and costly to switch to another provider.
- ➔ The risk of the university losing control over critical research datasets.

In conclusion, the scholar community fully acknowledges the importance of robust data practices but the supporting infrastructure and institutional culture are still not catching up. The ease of translation between the hypothetical best-case RDM frameworks and the day-to-day practice continues to create a significant drag on research productivity and collaboration. As the rallies to open, searchable and reusable data are becoming louder, universities and research centers must address the technical and human barriers to good data stewardship. This is the problem we are focusing on in this thesis and it will contribute to the larger debate concerning the way we can transform our institutions so that they can manage data in a more efficient and ethical manner.

In this project, we are delving into the development of a centralized data repository for the NOVA IMS University that will help solve the problems faced by both students and staff

helping improve student's productivity and decrease time loss. This data repository will provide students with a place to store datasets so that they can be used in the future by other students while giving each one the right access to improve data security and integrity.

2. Related Works

This chapter is dedicated to the exploration of some already existing data storage solutions in order to get an idea about the technologies used with their benefits and downsides. This can be used to help better contextualize our own system.

One of the three prominent data repository platforms are: CKAN, Zenodo, and Invenio

2.1 CKAN

CKAN or Comprehensive Knowledge Archive Network is a free to use open source data management system used for powering data hubs and data portals. It is essentially an open source open data portal for the storage and distribution of open data. It has become quite popular and a widely used solution mainly by governments in the implementation of their data portals. Some of the most notable countries adopting CKAN: U.S, Canada, Switzerland, Norway, Australia, Singapore, and Portugal.(Winn, J., 2013)(Flausino, M. G., Kozievitch, N. P., Fonseca, K. V., & Liu, E., 2024)(Umbrich, J., Neumaier, S., & Polleres, A., 2015)

Figure 1. *Porto Data Portal Powered by CKAN*

CKAN works as a modular, open source web based data management system that enables users to publish, search, share, and manage datasets. With being open source, it has a powerful ecosystem of extensions(or plugins) maintained regularly by community members which make it very expendable and highly customizable.

It is built using other open source technologies such as

- Python for the backend(handling server side logic, dataset management, access control, and API processing with libraries such as Flask and SQLAlchemy) but also the frontend using python libraries such as Jinja2 to render HTML pages
- PostgreSQL as a database management system, and Redis as a key-value store to support background jobs, session, coaching, etc.
- Solr as a search engine for searching through the data portals

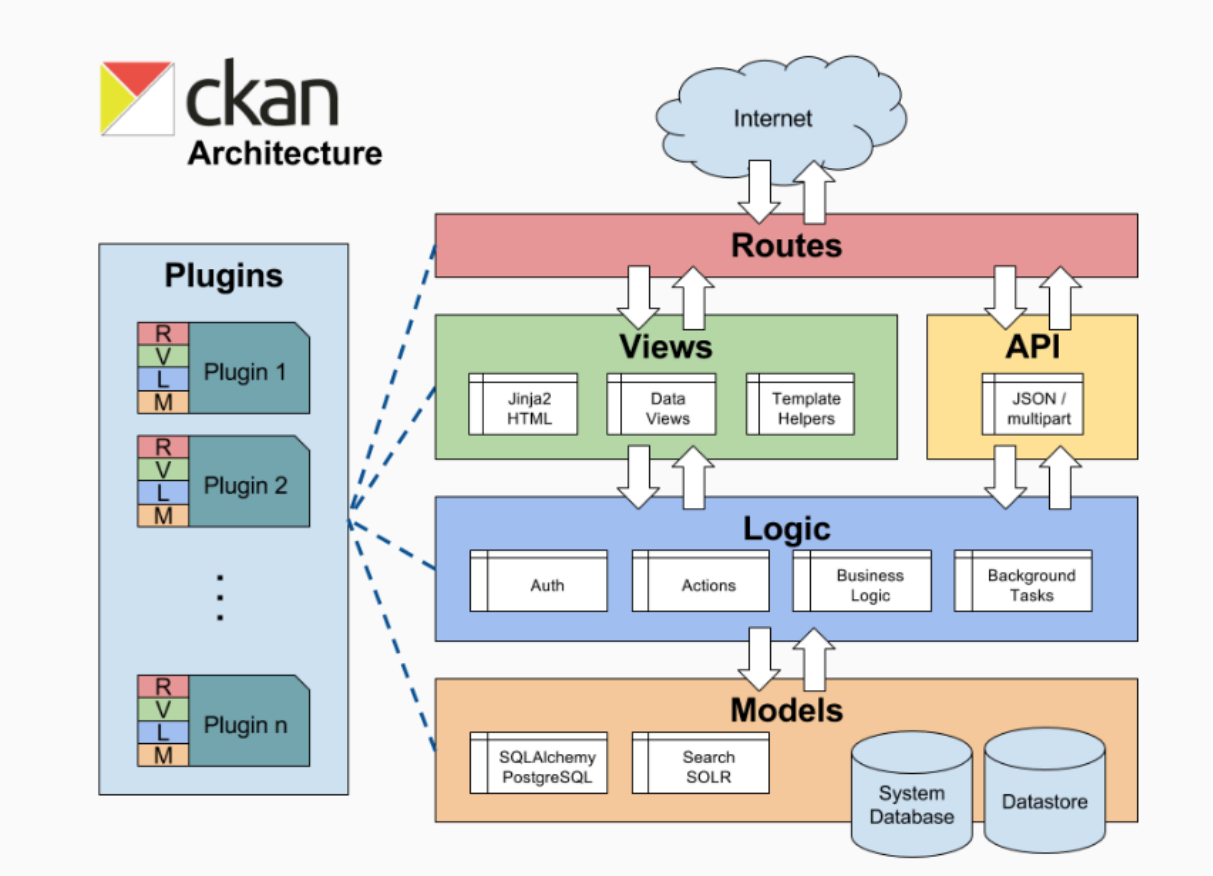


Figure 2. *: CKAN code Architecture*

The figure 2 showcases CKAN's code architecture with different layers and components that work in tandem with each other. The layers are as follow:

- **Routes**: They define the connection between CKAN URLs and either views or backend API calls that process requests and provide responses.
- **Views** (Frontend Web Interface): They process requests by reading data with an action function and return a response by rendering Jinja2 templates.

- **API** (Machine Interface): They also process requests. The main difference is that instead of interacting directly, we use scripts or third party tools and typically return JSON.
- **Logic**: It includes 4 different functions:
  - Auth: Handles permission and role checks
  - Action: Main logic handlers that serve to the main CRUD functions
  - Business Logic: Enforces rules like name unicity for example.
  - Background Tasks: Handles asynchronous operations such as indexing
- **Models**: Here is the database management system(SQLAlchemy + PostgreSQL) and the search tools (SOLR).
- **"Plugins"**: Although not a proper layer in CKAN per say but can be added for more features and benefits.

CKAN	
Pros	**Cons**
Robust metadata handling	Steep learning curve
Powerful API	Outdate frontend UX
Plugin architecture	Limited real-time features
Widely adopted	Complex plugin development
Integrated data storage options	Not tailored for research workflows
Open source and customizable	Requires external services

Table 1. *: CKAN Pros and Cons*

2.2 Zenodo

Zenodo is a well-known open-access research data infrastructure, which is constructed by CERN within the OpenAIRE project. The platform assists researchers of any field to store, share and cite any type of research product- be it datasets, software, publications, presentations or other digital research assets. The ease to use is one of the greatest strengths of Zenodo: you can upload files without any technical preparation and the system will automatically create a Digital Object Identifier (DOI) that will allow your work to be correctly cited and traced in the long run. Zenodo is under the hood an open-source Invenio framework, which implies that it is capable of receiving rich metadata, connecting with researcher identifiers such as ORCID, and facilitating open data sharing that adheres to the FAIR principles (Findable, Accessible, Interoperable, and Reusable). Due to its adherence to the principles of open science, Zenodo has become a preferred option in cases when an individual, rather than the entire institution, wants to share his or her work as soon as possible and with the highest possible degree of confidence. (Gurav, V., & Nagarkar, S. R., 2025)(Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. I., 2017)

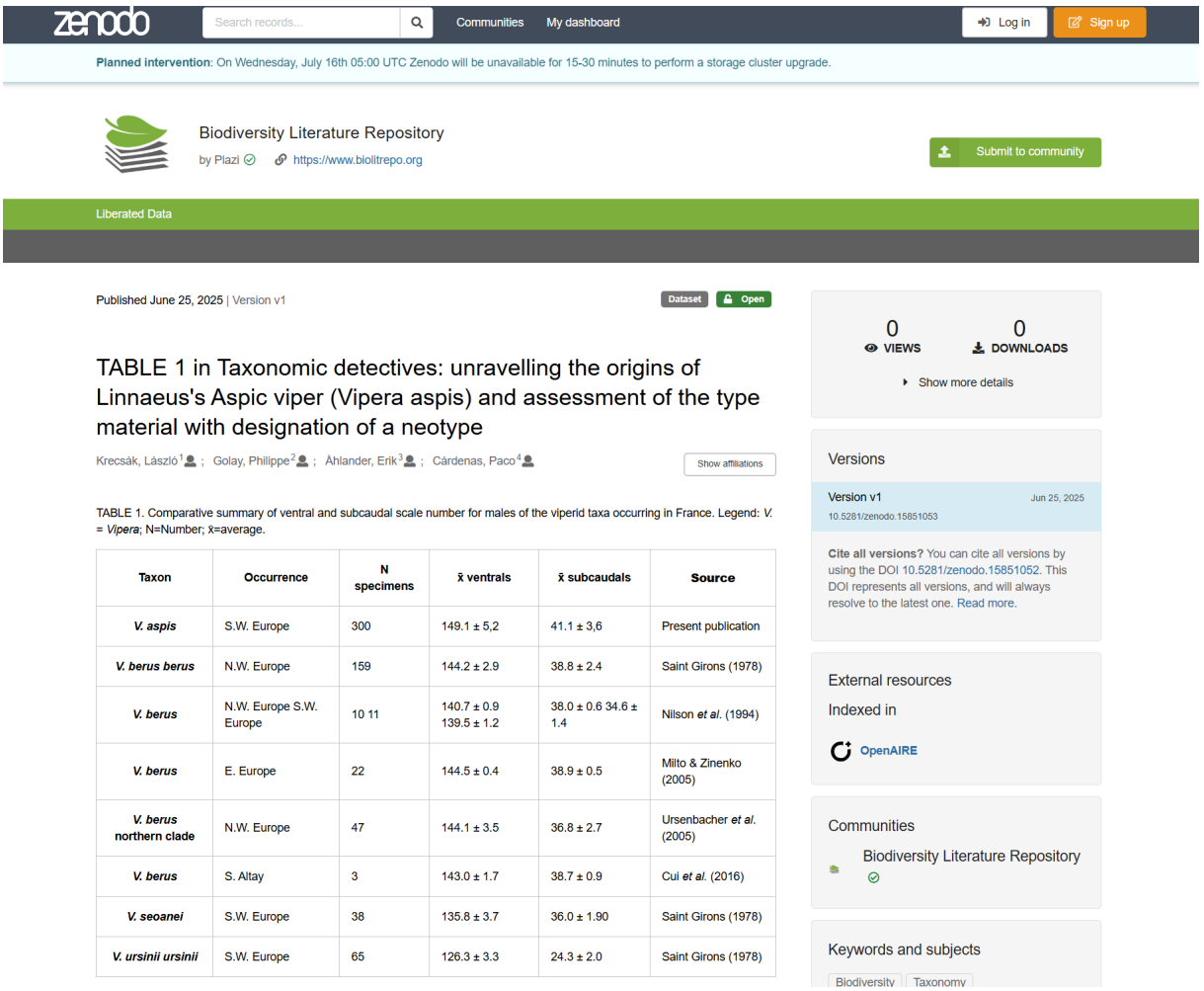


Figure 3. *: UI of the Zenodo data repository*

Zenodo is powered by the mature Invenio framework, a modular open-source initiative that allows it to scale and remain modular. Zenodo uses Python and Flask to process web requests and provide REST APIs to every aspect of metadata editing as well as file management. PostgreSQL stores structured metadata and user data under the hood and Elasticsearch does full-text search and filtering. Invenio-Files-Rest does the real work of uploading files, and the information is stored in the CERN EOS system or any other S3-compatible storage tank. Background tasks, such as indexing, assigning DOIs, or checking file integrity, are offloaded to a Celery and Redis queue system, meaning the frontend can remain responsive even as workloads accumulate. The user logins and permissions are OAuth2-based, so there is no need to fumble about plugging in to ORCID and GitHub. To make things reasonable, metadata validation and record serialization are executed on JSON Schema, which allows the researchers to extend the schema to the specific areas. Combined, this modular arrangement provides unique identifiers that do not change over time, social organization of content, and permanent access to scientific products, which is why Zenodo is one of the longest established and most technologically well-developed platforms in open science infrastructure. (Turner, G., Culhane, M., Delwar, I., & Sizemore, J., 2017)

Zenodo	
Pros	**Cons**
Free and hosted	Limited Customization
DOI assignment	Upload size limit
ORCID and Github integration	No fine grained access control
Supports datasets, software, and publications	Not ideal for sensitive/private data
User friendly UI	No institution wide management tools
Built on Invenio	Difficult to self-host

Table 2.*: Zenodo Pros and Cons*

2.3 Microsoft Dataverse

Microsoft Dataverse is a cloud based solution that easily structures a variety of data and business logic to support interconnected applications and processes in a secure and compliant manner. It is designed to be a central data repository for business data and powers many Microsoft Dynamics 365 solutions like Field Service, Marketing, Customer Service, and Sales. It is also part of Power Apps and Power Automate with native connectivity built right in. (Mora, L. T., 2023)

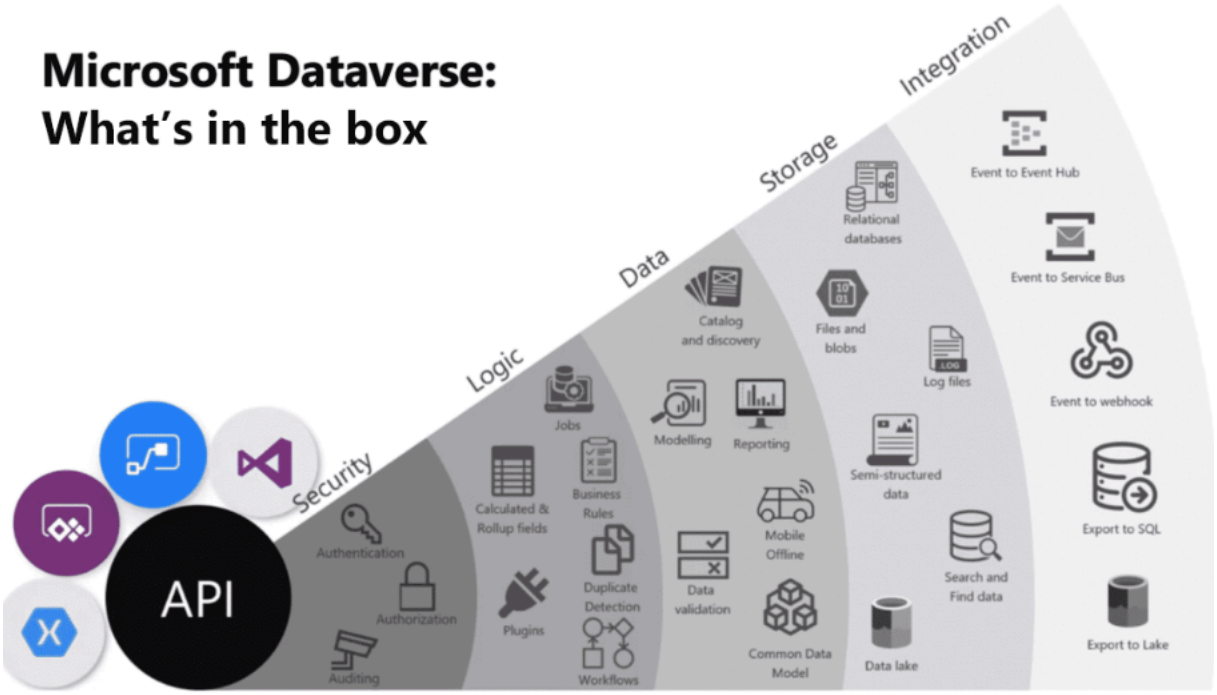


Figure 4 *: Microsoft Dataverse offerings*

Figure 4 showcases the many offerings of Microsoft Dataverse that would be explained down below:

- **Security**: Dataverse handles authentication with Azure Active Directory(Azure AD) to allow for conditional access and multi-factor authentication. It supports authorization down to the row and column level and provides rich auditing capabilities.
- **Logic**: Dataverse allows you to easily apply business logic at the data level. Rules are applied no matter how the user is interacting with the data and they can range from duplicate detection to business roles and more.
- **Data**: Dataverse provides the user with total control over the shaping of the data that allows for discovery, modelling, validation, and reporting of said data. This control ensures that the data looks how you wanted regardless of how it is used.

- **Storage**: Dataverse stores data in the Azure cloud, thus removing the burden of worrying about the whereabouts and scalability of this data. (Asundi, S., & Piridi, S., 2024)
- **Integration**: Dataverse connects in different ways to support your business needs. APIs, webhooks, eventing, and data exports give the flexibility to get data in and out.

Microsoft Dataverse is simply a cloud-based data platform, which operates on Microsoft Azure. It is designed to store both the structured and unstructured business data in a secure and manageable manner. Under the hood it has Azure SQL and Azure Blob Storage, so it is highly scalable and high-performance with relational data and file-based content respectively. Within Dataverse, we work with tables that can be customized (previously known as entities) that have rich metadata, relationships, and business rules. There is also role-based security and fine-grained access control to the table, row, and column, which actually assists in governance and compliance. The interface to Dataverse is modern and based on RESTful Web API (OData v4) and a variety of SDKs, which enables seamless development and integration with the other Power Platform tools, such as Power Apps, Power Automate, and Power BI. Besides that, Dataverse can give us a possibility to run business logic using out-of-the-box workflows, calculated fields, and plugins, thus we do not need to use external infrastructure. Throw in that it has native integration with Dynamics 365 and connectors to services such as Azure Synapse Analytics and Data Lake Storage, and you have a centralized enterprise-level data backbone that would be ideal in a modern low-code/no-code application development environment and analytics.

Microsoft Dataverse	
Pros	**Cons**
Deep power Platform Integration	License cost
Enterprise grade security	Limited customization
Scalable and cloud native	Proprietary ecosystem
Rich data modeling	Developer constraints
Low code support	Limited offline access
Automatic DOI style GUIDs and data versioning	Less transparency

Table 3.*: Microsoft Dataverse Pros and Cons*

2.4 Summary

The table below contains a summary of the difference between the previously mentioned solutions:

Feature	CKAN	Microsoft Dataverse	Zenodo
Hosting	Self-hosted(open source)	Cloud-based (Microsoft Azure)	Hosted by CERN (open access)
Tech Stack	Python(Flask), PostgreSQL, Solr	Azure SQL, REST API(OData), Power Platform	Invenio(Python), PostgreSQL, Elasticsearch
Use Case Focus	Government/open data portals	Enterprise/business app backend	Research data publication
Data Model	Custom datasets and resources with metadata	Tables with fields and relationships	Flat file uploads with basic metadata
Customization	Highly extensible via plugins	Limited; config via Power Platforms tools	Minimal customization
User Interface	Template-based, less dynamic	Integrated with Power Apps UI	Clean, simple web interface
Access Control	Role-based, per organization	Field-level security, RBAC	Limited(mostly public/open)
DOI Support	Optional via plugins	No native DOI support	Built-in DOI minting via DataCite
Versioning	Manual/version plugins	Table-level version tracking	File/dataset-level versioning
Best For	Institutional or national data platforms	Internal enterprise or institutional data apps	Sharing research datasets and code
License Model	Open-source(AGPL)	Commercial (included in Microsoft ecosystem)	Free, open access

Table 4 *: Key differences between CKAN, Microsoft Dataverse, and Zenodo*

The existing open-data services CKAN, Zenodo, Microsoft Dataverse are all ideal candidates during a large-scale deployment, but each service has its own trade-offs. CKAN and Zenodo are very capable but have the drawbacks of server-rendered UI, strict extension ecosystems and poor institutional adaptability. Microsoft Dataverse is a rich feature environment that is limited by a proprietary cloud environment that does not allow end users to customize the backend or integrate with third parties. Comparatively, the offered solution implements a light-weight, current stack: JavaScript (Next.js) on the frontend and MongoDB as the main database. This choice benefits from the flexibility of JavaScript in making fluid and responsive UI with real time interactivity thanks to the cache storage and React hooks, it also allows the app to be highly customizable to meet academic and students needs and workflows.

The issues that were identified in the course of user research and prototype testing and that were addressed by the implementation of the platform refer to such persistent pain points as the ambiguity of user roles, poor searchability, and the inability to support multiple versions of data at a time.

3. Design Thinking

Design Thinking is a human-centered approach to problem-solving; it keeps the user at the center of each of the development stages. It is a non linear and iterative process that helps foster and grow innovation by allowing us to better understand the users and the issues they face on a daily basis, thus helping us define the actual problems and needs that our project needs to solve. It differentiates itself from other methods by having its main focus being about making solutions that are appropriate for people as well as recognizing that everything is a system; you have to look and understand the nature of the system because you can't just solve little pieces.

It is best used when tackling unknown problems or what is called wicked problems. A wicked problem is a problem that is unique and where standard solutions don't work. That makes them trickier to work with as you have to approach them individually with a clean mindset. They also demand outside of the box thinking, immediate action and constant iteration which are essential landmarks of the design thinking approach as it offers practical and adequate methods and tools. (Razzouk, R., & Shute, V., 2012)(Dorst, K., 2011)

The Design Thinking has three main objectives in mind (which can be shown in more details in the figure 5): (Waidelich, L., Richter, A., Kölmel, B., & Bulander, R., 2018)

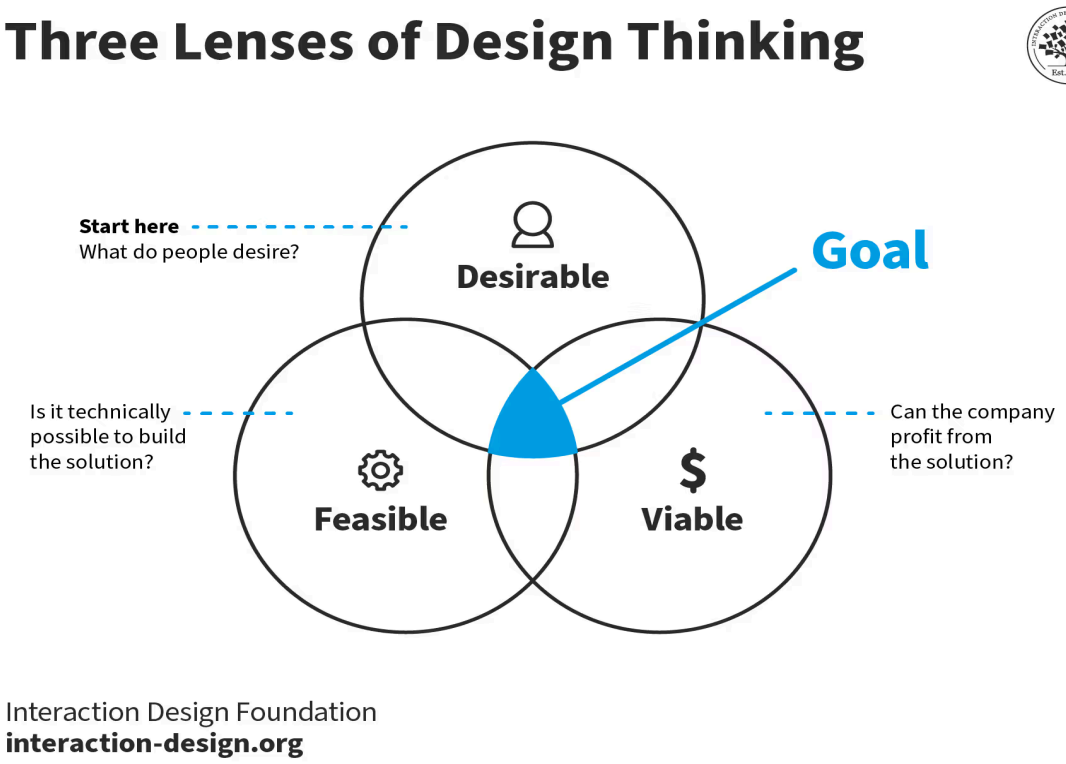


Figure 5 *: The Three Lenses of Design Thinking*

- **Desirability** (what do people desire?)**:** First and most important lens in the process, it looks at the needs and requirements of the people (user). It requires listening with an

empathetic ear to understand what the users truly need and not what the corporate think they want/need. It then leads to solutions that actually satisfy the end users point of view.

- **Feasibility** (is it technically possible to build the solution?)**:** Having properly determined the user needs and by extension the proper solutions, now comes the time to determine whether those solutions are feasible or not. Yes, everything is possible given all the time and resources in the world, but we have to evaluate if the solution is worth pursuing with our current resources. But, with that being said, we shouldn't be bothered too much with the technical restrictions at the first stages of the design thinking or else it will limit the capacity for innovation and creative solutions.
- **Viability** (can the company profit from the solution?)**:** Desirability and feasibility alone aren't enough, a company/organization needs to benefit from the final product one or another (either financially, or the problems it resolved are worth more than the invested cost). Here comes the viability lens to help stakeholders make their decision.

One of the pioneers of the design thinking approach is Stanford University's Hasso Plattner Institute of Design, their approach is considered by many as the “correct” way of the Design Thinking methodology. This process consists of five steps that are not necessarily sequential, as said before the design thinking method is non-linear, so some back and forth is needed in order to refine the solutions and constant feedback helps improve the first set of proposed solutions.

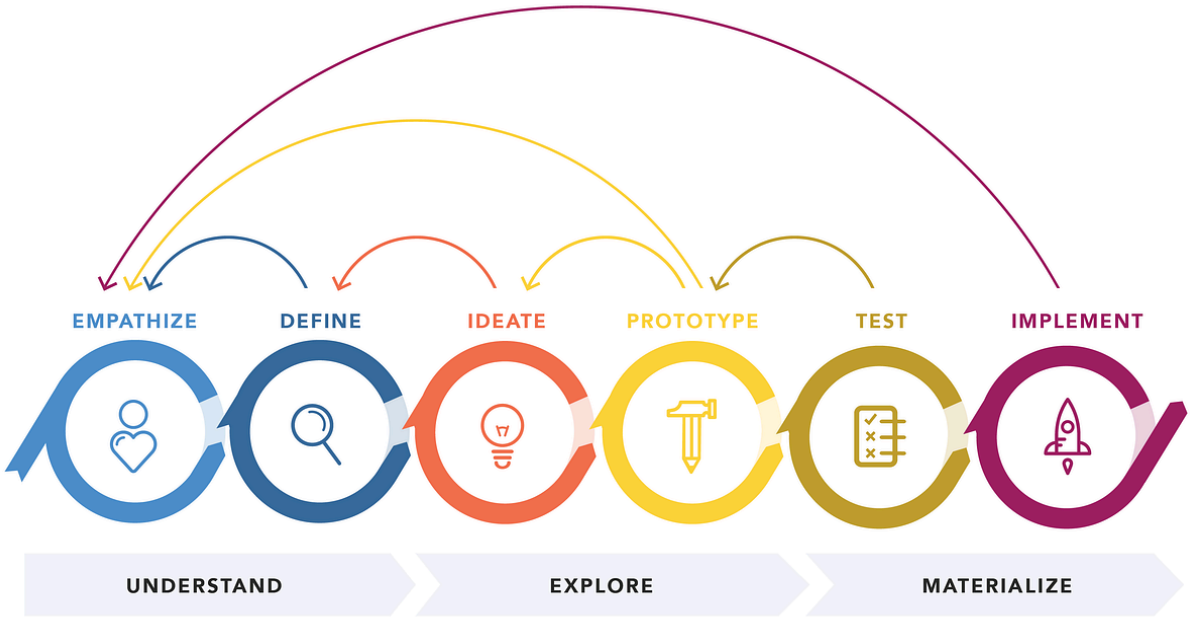


Figure 6*: Five Steps of Design Thinking*

These steps are as follow:

- **Empathize**: In this stage, the main objective would be for us to get a deep understanding of the users and the context they operate around. This can be done by direct interactions with the users where we can determine what problems they have that need solving. This direct involvement creates a sense of empathy and we can build that through observations, interviews, and also surveys that show what the user does, thinks and feels. Establishing empathy guarantees that the developed solution is grounded in real, lived experiences and answers to real life problems, not just some hypothetical requirements.
- **Define**: During this stage, we take the insights that we gathered during the empathize phase, we then analyze and synthesize them into clear, concise, and actionable problem statements. These statements represent the main issues facing the users and formulate them in a way that is easy to work with and lead to precise solutions. Here is not the time to generate ideas and solutions, it's only here to define key points and ensure alignment with what problem needs solving. A strong definition must balance between 2 main points specificity and flexibility; it must be elaborated in such a way where everyone working on the project has an idea on the end goal that needs working towards without defining all the necessary steps to get there, all creativity and spice is welcome; as long as it helps lead efficiently to the goal.
- **Ideate**: This phase leverages the previous phase; we take the problems defined earlier and we start by brainstorming potential solutions without immediate filtering. Some tools like mind mapping, sketching, or group workshops can help jump start ideas. That doesn't mean that all ideas will be developed but the more diverse the input is the better; it helps view the problem through different perspectives and encompassing all its aspects while at the same time encouraging an outside the box way of thinking and looking beyond the conventional and obvious answers. In the Ideate phase, the foundation for designing effective and user-centered solutions can be built.
- **Prototype**: This phase is pretty straightforward and self explanatory. We take on the suggested ideas from the previous step(Ideate) and we make them tangible. A prototype's main use is to be able to test on how the idea would perform in a real world setting and identify potential usability/technical issues.
- **Test**: Again, pretty self explanatory, we give prototypes to users to get genuine feedback. The main idea is to determine the usefulness, usability, and functionality of the prototype to the individuals whom it is designed for. It helps us get a sense of what works and what we need to improve upon. One important thing is that this step is not a one time thing, we then go back to the Prototype step with a new improved version that goes through testing once again until the users are satisfied with the end result.

The Alternative was the agile method that focuses on iterative development and the delivery of solutions in short cycles. The following table showcases the differences of these 2 methods:

Feature	***Design Thinking***	***Agile***
Primary Focus	❖ Problem definition ❖ Understanding user needs ❖ Ideation	❖ Iterative development ❖ Delivering working solutions
Goal	Discovering the right problem to solve	Solving the problem in an efficient and effective fashion
Approach	❖ Human-centered ❖ Empathetic ❖ Exploratory	❖ Time-boxed iterations(sprints) ❖ incremental delivery
Key Activities	❖ Empathize ❖ Define ❖ Ideate ❖ Prototype ❖ Test	❖ Planning ❖ Development ❖ Testing ❖ Review ❖ Retrospective
Best Suited For	❖ Exploring complex problems ❖ Identifying unmet user needs ❖ Innovation	❖ Projects with evolving requirements ❖ Rapid feedback loops ❖ Delivering value incrementally
Output	❖ Deep understanding of user needs ❖ Potential solutions ❖ Prototypes	❖ Working software/product ❖ Validated learning ❖ Continuous improvement
Emphasis	❖ Empathy ❖ Creativity ❖ User-centricity	❖ Speed ❖ Flexibility ❖ Collaboration

Table 5 *:Main Differences between the "Agile" and "Design Thinking" methods*

Although the Agile method sounds appealing as the expected output is a working software, the fact that the problem to solve is unknown plus the fact that it is not so user-centric helps tip the balance in favor of the Design Thinking.

With all the previous knowledge acquired, it becomes evident that the Design Thinking approach would be a good fit for this project. Our problem falls into the category of wicked problems stated above, research students suffer from the lack of data collaboration and from data loss and reuse, and available solutions are either costly or hard to use so they need a unique solution that adapts to their specific needs. Plus, the constant testing and feedback helps fine tune the solution to their specific needs and demands.

4. Requirements

Before developing anything, we first need to start by defining the requirements. But we should not simply gloss over this step, making sure that the proposed solution is as usable as it is meaningful to those it is intended for is quite the important matter. Shrinking the context to an academical environment that is the target of the scope of this project, we can find various users and concerned parties; on one hand, we find students and researchers and on the other hand administrative staff and IT personnel that both interact with data in different ways. This divergence of expectations requires the involvement of all the voices during the design process. By listening to what they directly need and the sloppy, implicit challenges hard-coded into their everyday workload, you can identify gaps that would not otherwise have been seen. Additionally, institutional stakeholders demand things such as, can this work conform to regulations or how easily can it be expanded as more jump on board. It is a matter of finesse to line everything up: the proper configuration must be adapted to each person and their workflow, but it also has to comply with the macro objectives of the entire institution. That balance is described in this chapter, with the core expectations, concerns, and priorities expressed by students and stakeholders outlined around which we could construct the final system.

4.1 Students

Firstly, we start by reviewing students as they are the main focus of this project; the data repository is made for them after all, so it only makes sense to first start by understanding their needs and expectations from such a project.

Two mutually complementary research designs have been utilized; a qualitative empathize phase, which is based on structured interviews, and a quantitative investigation phase that is constructed on a broader survey. The empathize stage involved semi-structured interviews that mostly targeted data science master students that probed into their direct experience on data quality and management. In tandem with the qualitative study, the quantitative stage initiated a survey that would help detect even more generalizable perceptions of the same issues: perception of, challenges around, and actual practices with data quality and management. A wider population was surveyed and therefore more comparisons and stronger conclusions could be made.

After having led those interviews and distributed the survey, now comes time to analyse them and gather insight. Those are main problems we discovered that the students face while working on their research:

- **Inconsistent and Inefficient Data Management Practices**: Students usually simply rely on ad hoc methods like naming files manually and unstructured folder hierarchies which can lead to confusion, mismanagement and errors. Also, the lack of

some standardized practices and guidelines for data storage cause problems when it comes to data consistency and ease of access.

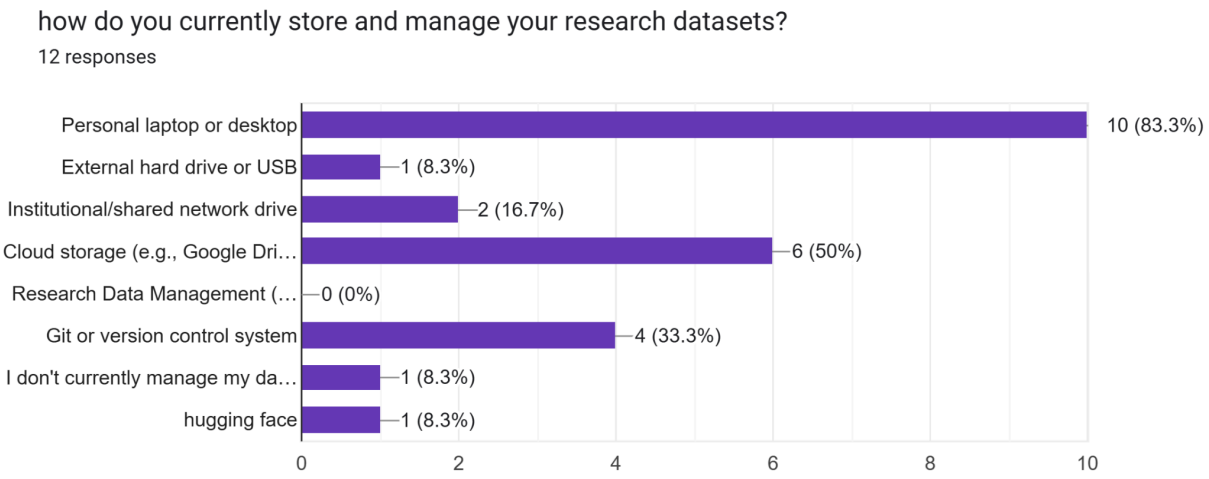


Figure 7 *: Bar Chart showing the most popular data management practices*

- **Poor or Nonexistent Version Control**: The most widely used technique among students is the manual renaming of the files (data_v1, data_final). This method is not robust at all, bloats the limited storage and increases the risk of confusion, data loss and duplication.

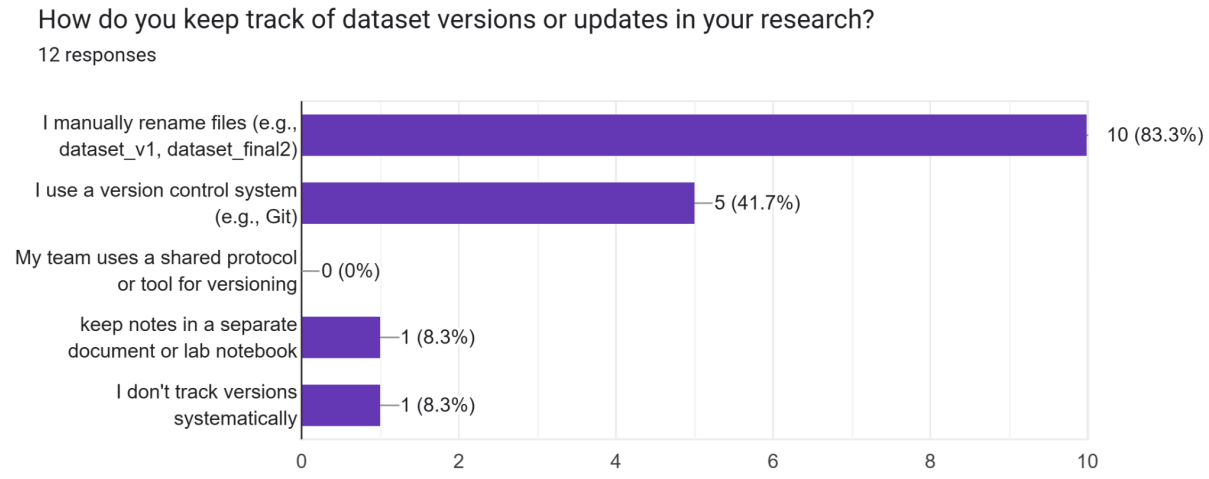


Figure 8 *: Bar Chart showing the most popular data "versioning" practices*

- **Limited or Inadequate Metadata and Documentation**: Data that is lacking in terms of information or proper labels can be quite annoying to work with, and even in some cases, lead to unnecessary repetition of effort or worse; a faulty analysis. During the interviews, the main culprit behind this lack of documentation/proper labeling was

the lack of motivation and time pressure. This causes problems and reduces the potential for the data to be used by another researcher.

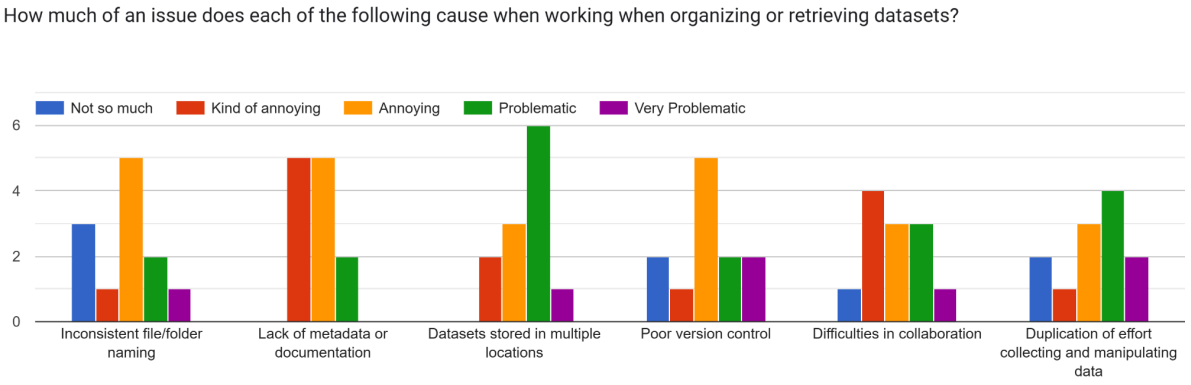


Figure 9*: Histogram determining the degree of annoyance each issue cause to the users*

- **Collaboration Challenges**: Students don't have the proper tools to be able to properly collaborate on certain projects which diminish the incentive to do so for some. Most of them rely mainly on email attachments that have a size limit or cloud storage option that offer additional space for a price.

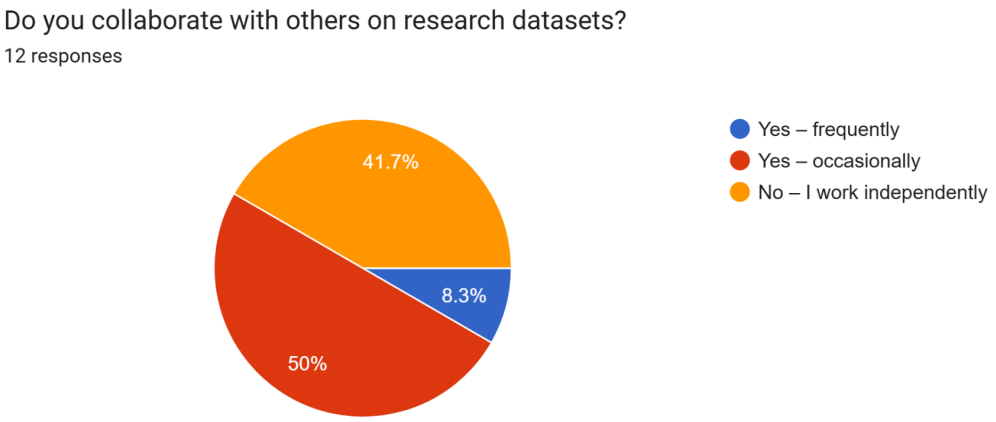


Figure 10 *: Pie chart showing the frequency of students collaborating during projects*

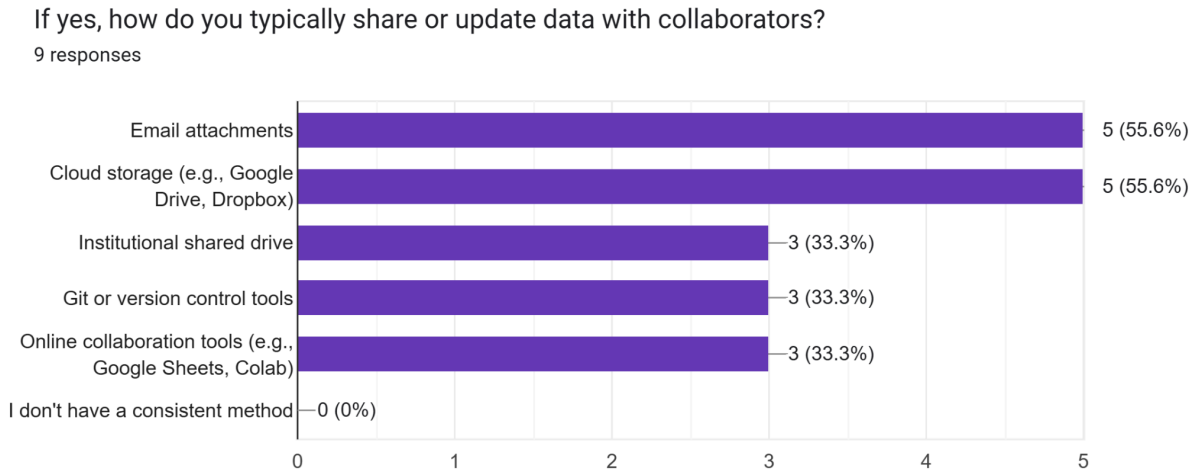


Figure 11*: Bar Chart Showing the popularity of data sharing and collaboration methods*

- **Difficulty Ensuring Data Privacy and Security**: Students were quite concerned about the security, privacy, and confidentiality while sharing data; you lose control over it which might lead to it being lost or even breached.

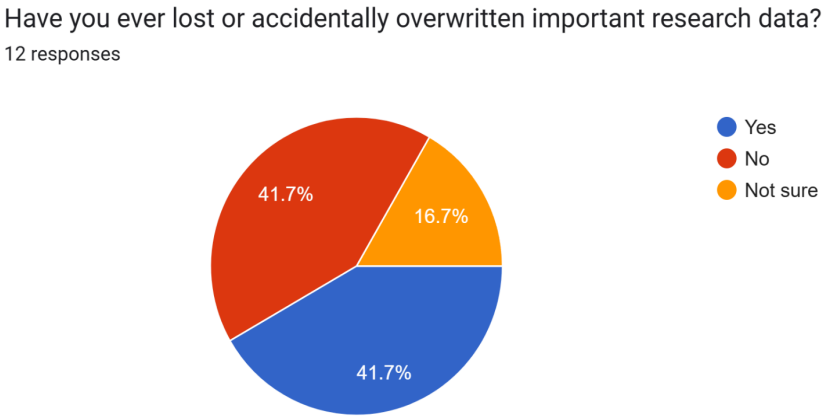


Figure 12*: Percentage of people that have lost data one way or another*

- **Technical Compatibility and User-Friendliness Issues**: The existing tools are not necessarily compatible with the types of data they work with and some of them are quite complex. Some students pointed out that if it would take them hours to be able to learn how to properly use the tool, why bother? they can just spend that time collecting data.

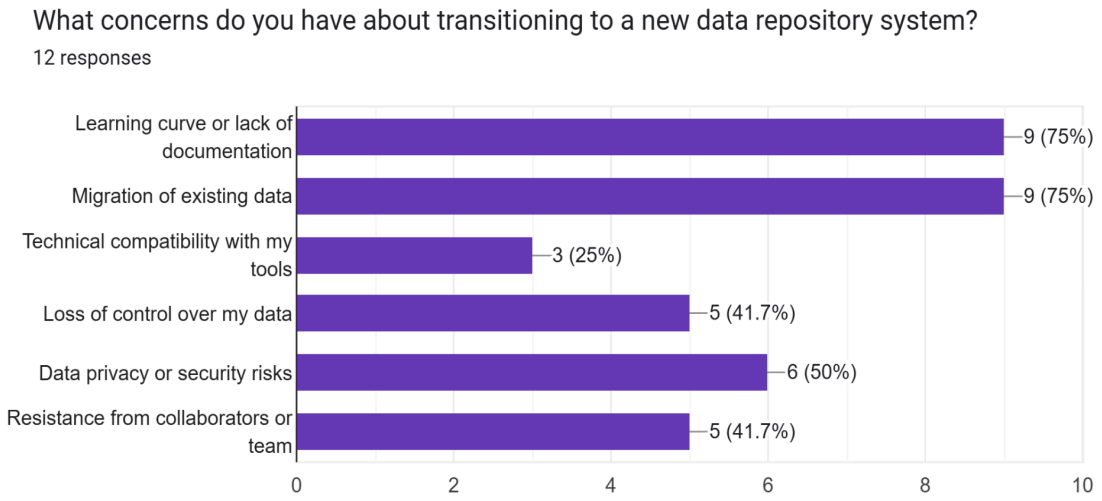


Figure 13*: Popularity of concerns about transitioning to new data repository systems*

So we can safely assume that students lack a systematic, secure, and user-friendly methods for managing and collaborating on research datasets, leading to significant inefficiencies, reduced data quality, and heightened risks of errors or data loss. They need an intuitive, integrated solution that automates documentation, standardizes version control, ensures security and privacy, and simplifies collaborative data management.

The following figure 14 shows the most important features for a data repository in the students eyes.

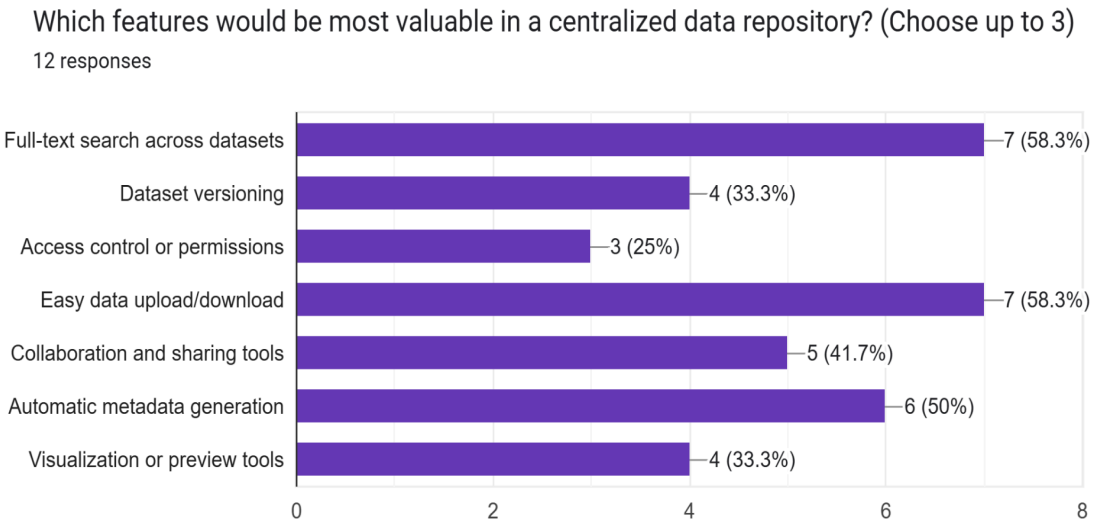


Figure 14*.: Importance of feature that should be available in a data repository*

4.2 Stakeholders

After having noted students' needs and expectations from such a project, we should also address and understand the needs of our stakeholders as well. Their opinion is important for the successful implementation of any solution, especially in an academic environment. Effective and permanent communication between the stakeholders and the developer side is quite essential so the designed solution would be properly aligned with the universities practical, operational, and strategic goals. Getting to understand stakeholders' concerns and requirements helps the developers determine a solution that is as technically suitable as financially sustainable and feasible under the university's constraints. Including their input from the beginning is a good thing as it helps with reducing resistance to change and would improve adoption rate and ensure long term success and effectiveness of the chosen solution.

In this project, for understanding stakeholders expectations, I opted for regular meetings with my supervising professor; Flavio Luis Portas Pinheiro; as it was an accessible and efficient way for more detailed back and forths. These meetings not only served to discuss progress and gather feedback but also to clarify the staff requirements and make sure what I was developing was what they actually needed and expected. This approach based on a mix of formal and informal communication made everything quite clear and that the objectives were respected.
The stakeholders had expectations and needed the application to be:

- Secure
- Compatible with existing university IT infrastructure
- Able to support multiple file formats
- Easy to use
- Customizable and scalable
- Able to track and monitor students participation and content quality
- Financially feasible

4.3 Summary

After gathering both the students and stakeholders' needs and concerns for this project, we can start answering by proposing a solution that would take all these issues into consideration. The stakeholders want the solution to be financially feasible and keep control over their data so a cloud solution like Microsoft Dataverse for example would not cut it. Prices can grow exponentially with the growth and adoption of the app plus there is also the risk of vendor lock-in that can lead to a loss of data ownership.
They want the application to be easy to use and customizable, so I opted for Next JS as it can create a highly customizable and clear UI greatly simplifying the user experience. It also has

the benefit of also being able to solve backend problems making it a very viable and strong full stack solution. (Patel, V., 2023)(Dinku, Z., 2022)
They also needed the solution to be scalable and able to support multiple file formats while also keeping the cost to a minimum so the choice of MongoDB for the database was quite obvious (Krause, M., 2024)(Győrödi, C., Győrödi, R., Pecherle, G., & Olah, A., 2015)(Chauhan, A., 2019). It is also easy to use and implement with Next JS.
For security purposes, I opted for a role based access to be implemented for feasibility and technical reasons.Having a complex system that guarantee premium security consist in a thesis topic in of itself so I opted for a more simple but sufficient approach (Yang, J. J., Li, J. Q., & Niu, Y., 2015). Two roles, one for each side(students and staff). Students will just have access to the necessary features like uploading and exploring datasets while only being able to edit their own entries. As for the staff role, they will have full access to everything even editing and deleting datasets deemed low quality.
For monitoring and tracking student participation, I added a voting system; users can either upvote or downvote a dataset entry and the download count will be shown to help determine which are the most popular databases for example. Also there will be an approval system where the staff can approve a dataset giving it more credibility.
P.S: The approval is not necessary for the visibility of the database, even “non approved” databases can be used and downloaded by users, it just adds a “quality” stamp.

5. App Framework

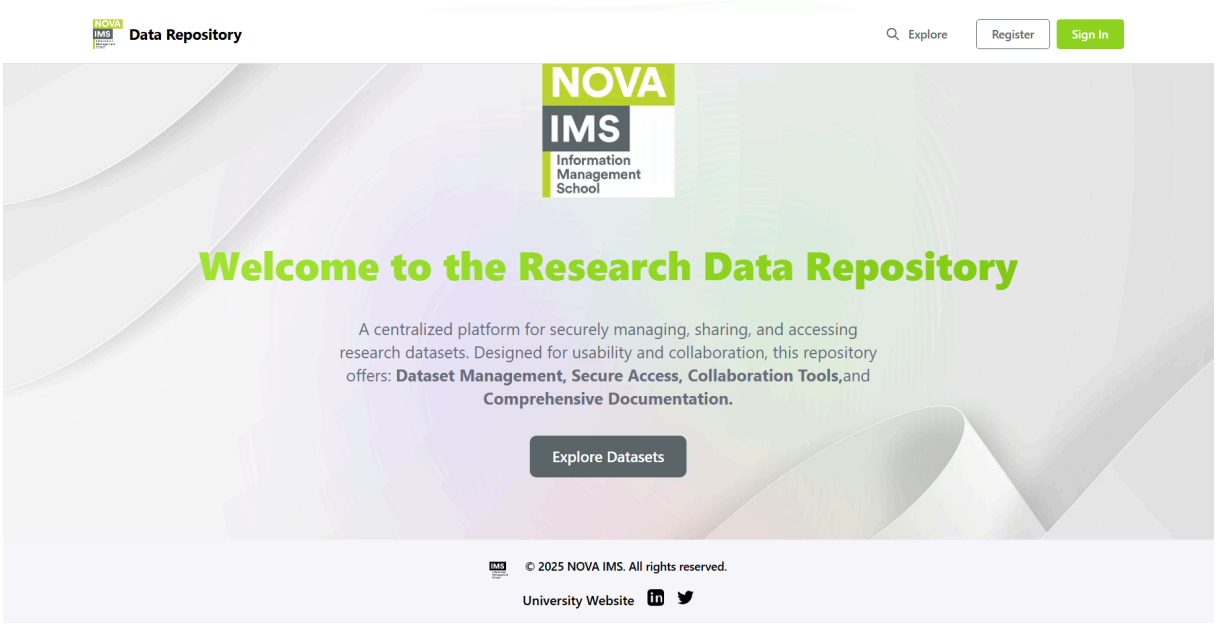


Figure 15*: App Home Page*

As mentioned previously, the data repository is created using NEXT JS that takes care of the frontend and backend thanks to its combination of static and server-side rendering capabilities for performance and SEO. The application is structured around API routes in the /pages/api directory to handle backend logic such as authentication, CRUD operations, and business logic. These API routes connect to a MongoDB database using the official mongodb Node.js driver or Mongoose as an ODM for schema modeling. A utility module that preferably uses a singleton pattern in order to ensure a single database connection instance across API requests manages the connection to mongodb. The frontend greatly benefits from the reusability and the adaptability of react components as well as other features such as data fetched with getStaticProps, getServerSideProps. The app uses a modular folder structure that helps with the principle of separation of concern, each part is stored in a proper folder(components, services, utilities, and API logic). For the authentication and session handling, NextAuth.js takes care of that with the benefit of being easy to integrate with mongoDb for the storage of sessions and user data.

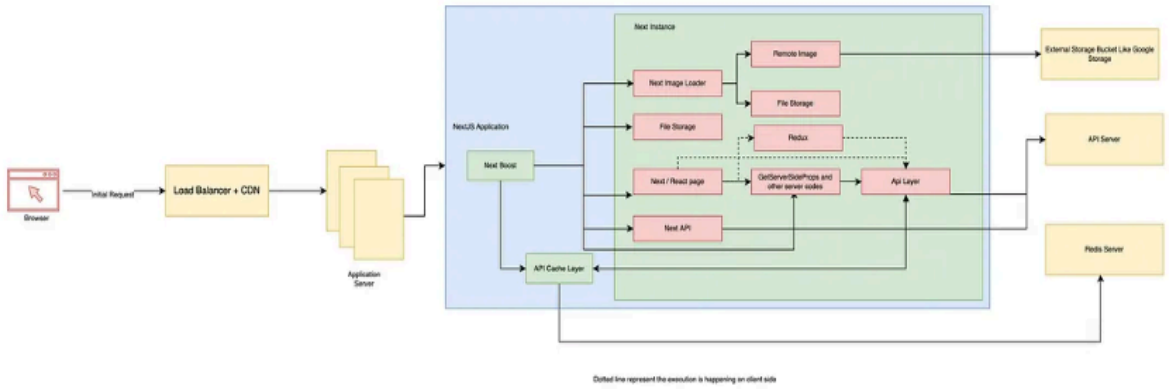


Figure 16 *: Architecture of a scalable Next JS application*

Figure 16 presents how it looks like for a scalable Next JS application with various components that handle requests, render pages, manage APIs, and optimize assets.
It all starts when a user initiates a request from the browser. The load balance and CDN(Content Delivery Network) start by caching static content and distributing traffic to improve performance. The request is then sent to the application server that forwards it directly to the core Next.js layer. Next Boost improves performance, and the API cache layer reduces the number of duplicate API requests. Every instance of Next.js does page rendering, either statically or on the server-side, and communicates with the next API that implements back-end logic. Next Image Loader and file storage components, take care of handling assets by fetching them either from a local source or an external one (Google Cloud for example). It also offers the possibility of being able to work with external services like dedicated API services. Adding to that, it also supports client side data fetching and execution(dashed lines) that creates a balance between server side rendering and browser based interactivity. This architecture has the benefit of being performant and modular which is a good plus for every modern web application that needs to be flexible and scalable. (Srivastava, S., Shukla, H., Landge, N., Srivastava, A., & Jindal, D., 2024)(Krause, M., 2024)

Figure 17 *: UI of the visual difference between the 2 roles*

The data repository also features a role based access feature that consists of 2 different roles: user/student and admin/staff. The role staff gives access to everything the app has to offer; from uploading to updating and editing the dataset. Having access to the voting and also the approval system. But the user has only access to the main features such as exploring and uploading datasets, editing his own submissions and of course the voting system.

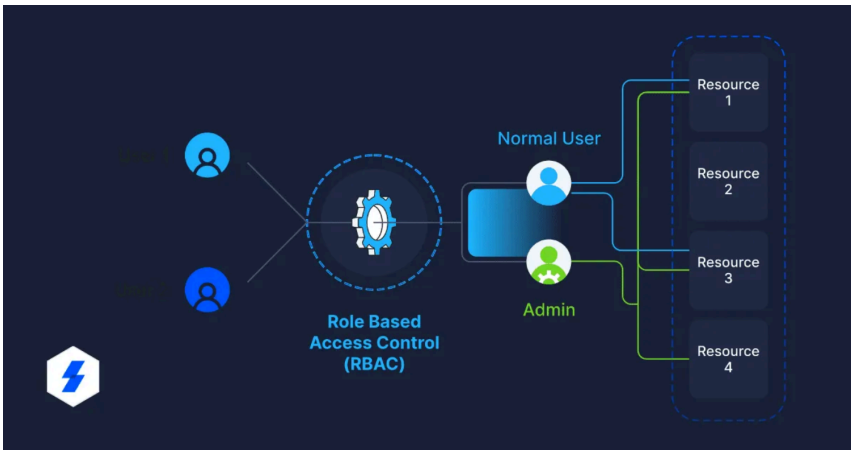


Figure 18*: Diagram explaining Role Based Access Control*

An important feature that helps the user navigate the datasets and make their life easier is the search and filter feature that helps you scurry through all datasets quickly and effectively.

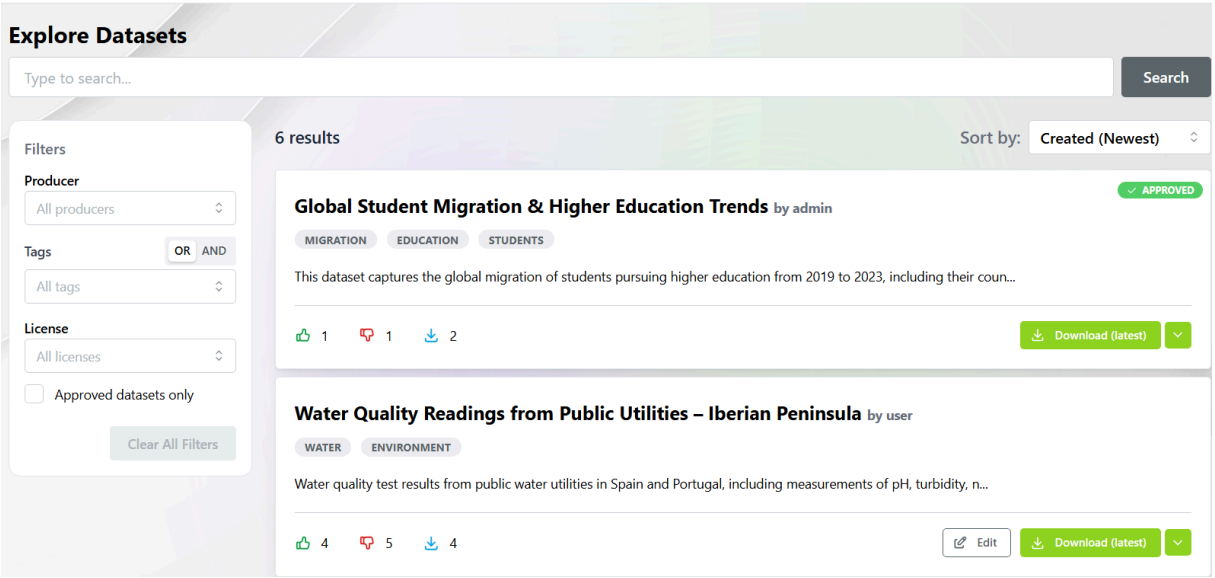


Figure 19 *: Search and Filter option in the application*

There is also a voting system, a download counter for each dataset, and an approval system that helps give more credibility to a dataset and help users make an easier decision in case there are multiple similar datasets.

Figure. : Upvote and Approval system

The process to upload data was kept easy and straightforward; you click on the add dataset on the top layer then you just fill up your metadata and you're good to go.The metadata should be solid though as it helps in other processes as well such as versioning. (Soules, C. A., & Goodson, G. R., 2003)

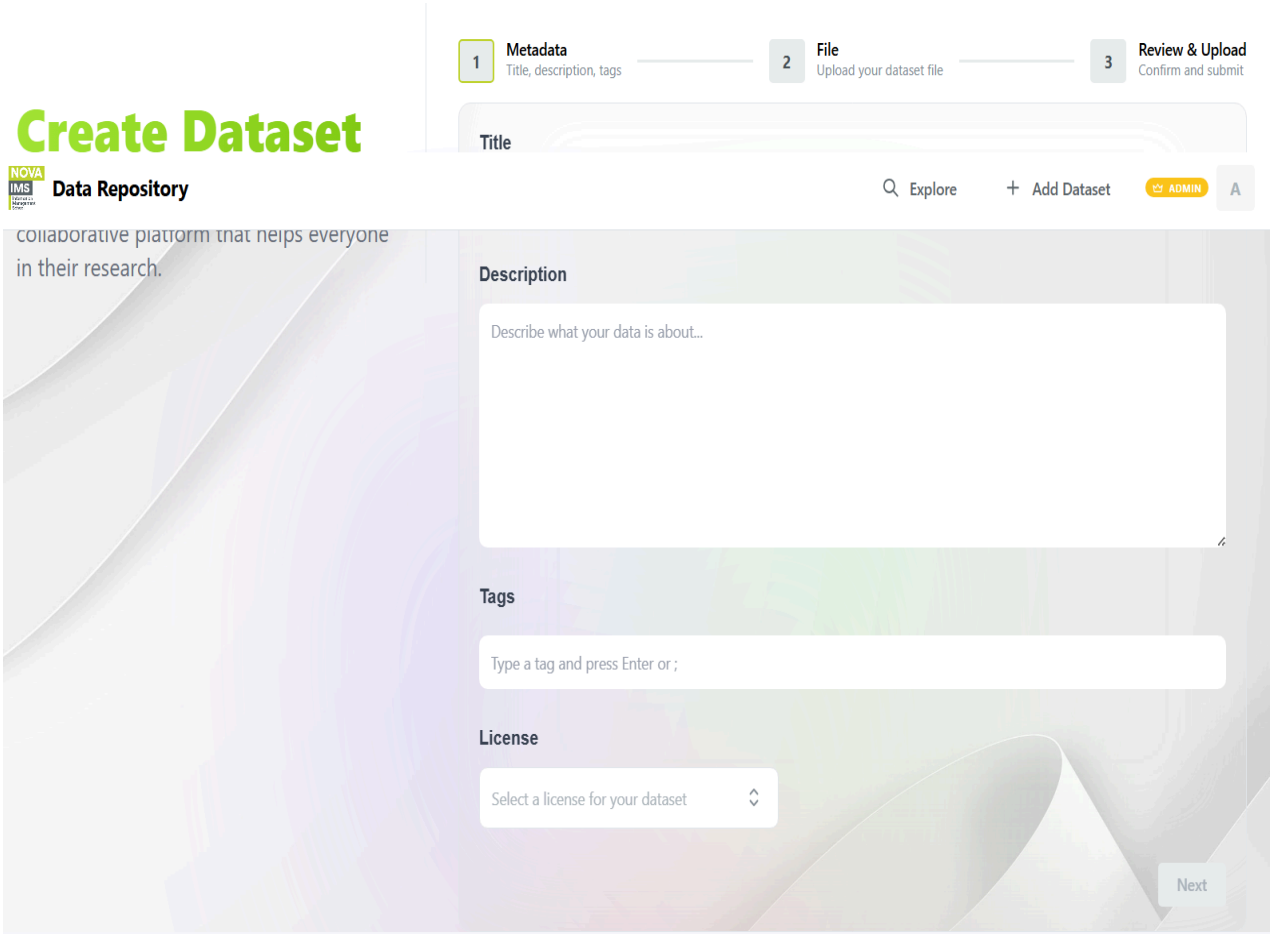


Figure 20 *: Create Dataset page*

You can also easily check and download other versions of a dataset.(Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R. R., & Asmi, A., 2021)

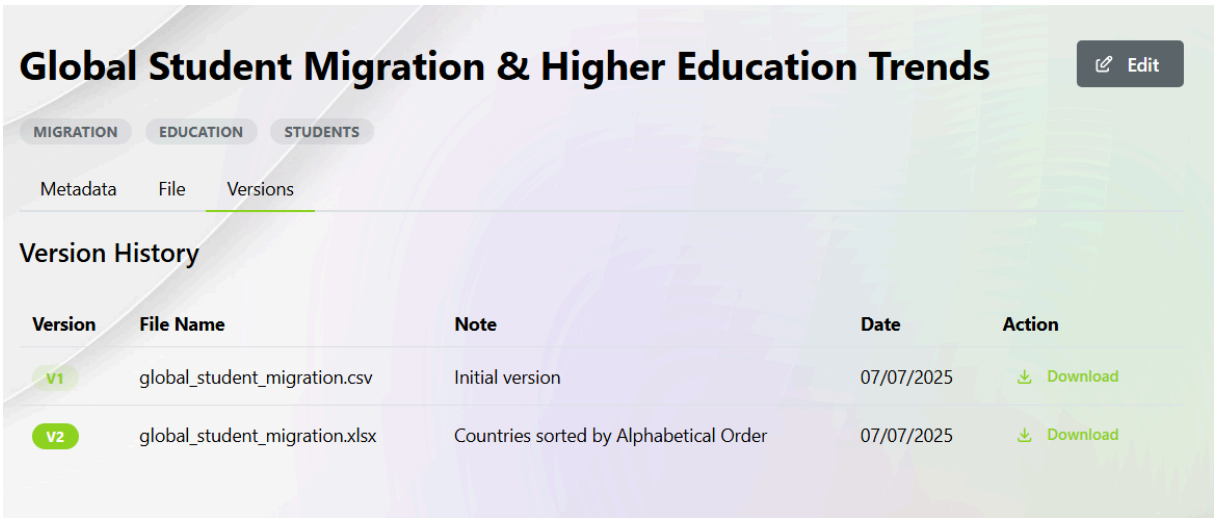


Figure 21 *: Version Panel*

After the app was fully developed, I asked some colleagues to test it and give me their feedback to see whether it indeed solves the core issues that were discussed previously.

I had one testing/feedback session with one of my classmates. During this session, one big positive that they mentioned was the UI, they liked the look and feel of it and said it was pretty to use. They also appreciated that the main functionalities are there(data storing, uploading, and exploring) plus the filtering and recherching feature gave good results.
However, one thing that caused the issue was the slow response time that was due to technical reasons; the app is made using the free version of mongo Atlas as a proof of concept and can be improved on with better providers. The true implementation will rely on the university´s infrastructure so that issue would be resolved. Another thing that was mentioned was the lack of visualisation and preview of the dataset before downloading it. I like the idea and it is considered as an added feature for future iterations.

6. Conclusions and Future Research

This project had the intention of addressing the rising need for having a central, concise, and easy-to-use data repository that suits academic research environments. In order to do this, we compiled the information based on interviews, surveys, and prototype testing. Based on such findings, we constructed the platform with particular reference to usability, transparency, and institutional relevance. The last system, which was constructed using a fullstack Next.js solution and MongoDB as a database, has the following functionality: upload/download dataset, manual metadata input, role-based access control, and versioning of datasets, which were designed in accordance with the daily operations of student researchers and institutional stakeholders.

This solution is more simplicistic and flexible and better integrated into the academic environment than other platforms such as CKAN, Zenodo, and Microsoft Dataverse, which have more complex user experience and are less suited to academic environments that display a high level of technical independence and flexibility.

Nevertheless, the development process also revealed the areas that should be explored in more detail. Future studies could improve search and filterable capabilities with more advanced natural language processing (NLP) to be able to retrieve more targeted datasets on the basis of descriptive metadata. Improved interoperability would be possible through the use of common standardised metadata export formats (such as JSON-LD or RDF) to enable the repository to connect to wider open science infrastructure. Real-time collaboration tools or data visualization embedded would also go a long way in making the experience as a researcher much more satisfying.

Lastly, more sustainable models in the medium- and long-term would make it relevant and impactful in the future, as well as more technically aligned with the FAIR data principles. These guidelines indicate the way towards the transformation of the platform into a platform that would encompass the entire institution in terms of data stewardship in academic research.

Bibliographical References

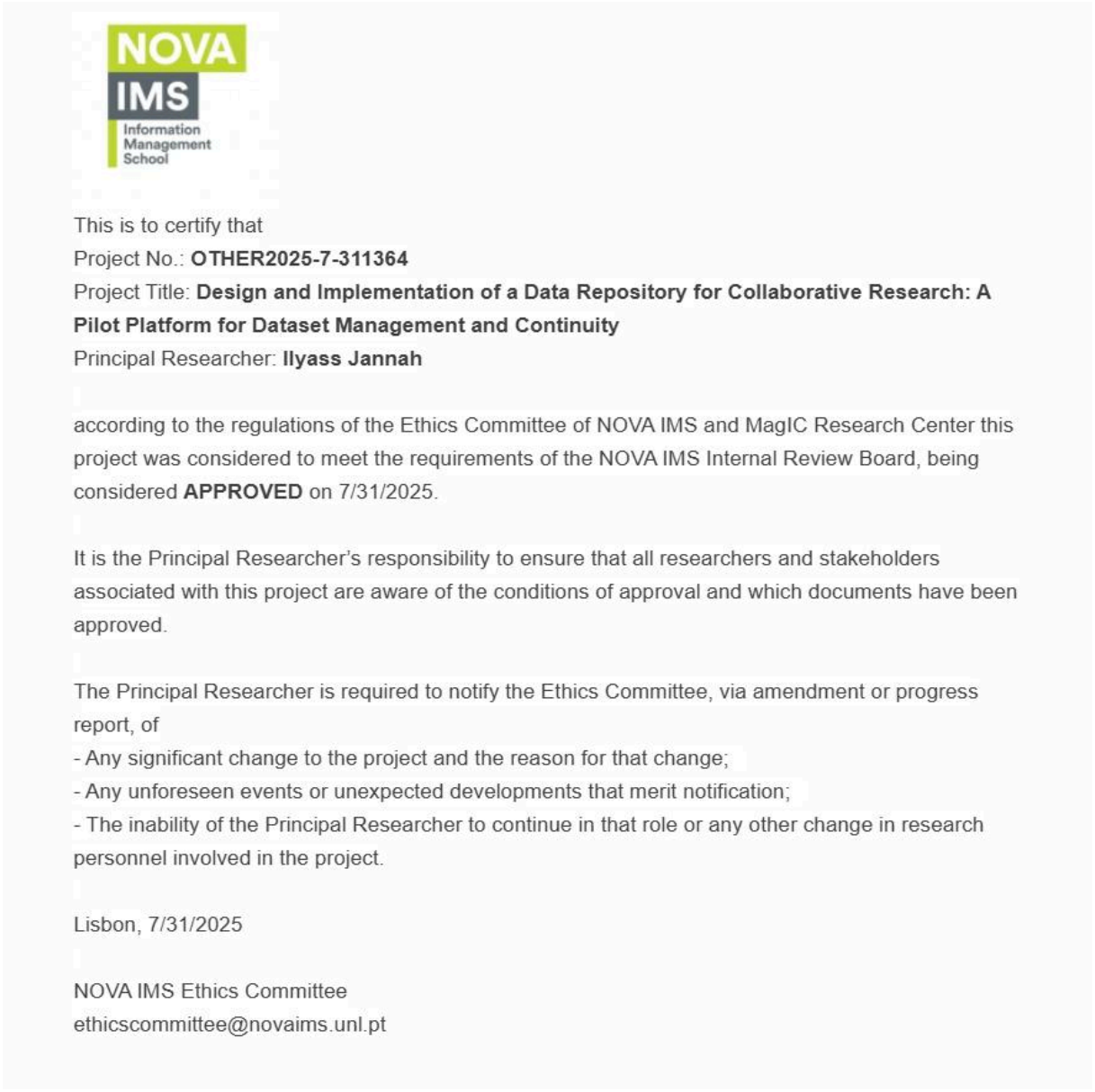
1. Almeida, F., & Calistru, C. (2013). The main challenges and issues of big data management. *International Journal of Research Studies in Computing*, *2*(1), 11-20.. https://doi.org/10.5861/ijrsc.2012.209
2. Asundi, S., & Piridi, S. (2024). Enhancing Role-Based Access Control (RBAC) in Dataverse for Secure Enterprise Applications–Analyzing Advanced Security Models and Governance for Power Apps and Dynamics 365 CRM. https://www.researchgate.net/profile/Satyanarayana-Asundi/publication/392064355_Enhancing_Role-Based_Access_Control_RBAC_in_Dataverse_for_Secure_Enterprise_Applications_-Analyzing_Advanced_Security_Models_and_Governance_for_Power_Apps_and_Dynamics_365_CRM/links/6831c9a2d1054b0207f24ed5/Enhancing-Role-Based-Access-Control-RBAC-in-Dataverse-for-Secure-Enterprise-Applications-Analyzing-Advanced-Security-Models-and-Governance-for-Power-Apps-and-Dynamics-365-CRM.pdf
3. Briney, K. A., Coates, H. L., & Goben, A. (2020). Foundational practices of research data management. https://scholarworks.indianapolis.iu.edu/items/9c299f1c-b2f9-4412-bdba-62a8eb1ebe35
4. Chauhan, A. (2019). A review on various aspects of mongodb databases. *International Journal of Engineering Research & Technology (IJERT)*, *8*(05), 90-92. https://www.academia.edu/download/60661268/a-review-on-various-aspects-of-mongodb-databases-IJERTV8IS05003120190921-10051-a8lu8n.pdf
5. Daki, H., El Hannani, A., Aqqal, A., Haidine, A., & Dahbi, A. (2017). Big Data management in smart grid: concepts, requirements and implementation. *Journal of Big Data*, *4*(1), 13..https://doi.org/10.1186/s40537-017-0070-y
6. Dinku, Z. (2022). React. js vs. Next. js. https://www.theseus.fi/bitstream/handle/10024/750122/Dinku_Zerihun.pdf
7. Dorst, K. (2011). The core of 'design thinking'and its application. *Design studies*, *32*(6), 521-532. https://doi.org/10.1016/j.destud.2011.07.006
8. Dunie, M. (2017). The importance of research data management: The value of electronic laboratory notebooks in the management of data integrity and data availability. *Information Services and Use*, *37*(3), 355-359.https://doi.org/10.3233/ISU-170843
9. Flausino, M. G., Kozievitch, N. P., Fonseca, K. V., & Liu, E. (2024, October). Open Data Portals-A case study of Challenges and Opportunities. In *Brazilian e-Science Workshop (BreSci)* (pp. 104-111). SBC. https://doi.org/10.5753/bresci.2024.244096
10. Gurav, V., & Nagarkar, S. R. (2025). ZENODO: A PLATFORM FOR OPEN ACCESS AND SUSTAINABLE DIGITAL RESEARCH REPOSITORY. *TechnoLibrarianship: A Gateway Towards Future Libraries-2025, Shivaji University, Kolhapur, PP-152-159*. https://doi.org/10.2139/ssrn.5246452

11. Győrödi, C., Győrödi, R., Pecherle, G., & Olah, A. (2015). A comparative study: MongoDB vs. MySQL. u: 2015 13th International Conference on Engineering of Modern Electric Systems (EMES). https://doi.org/10.1109/EMES.2015.7158433
12. Hulsen, T. (2020). Sharing is caring—data sharing initiatives in healthcare. *International journal of environmental research and public health*, *17*(9), 3046. https://doi.org/10.3390/ijerph17093046
13. Klump, J., Wyborn, L., Wu, M., Martin, J., Downs, R. R., & Asmi, A. (2021). Versioning data is about more than revisions: A conceptual framework and proposed principles. *Data Science Journal*, *20*, 12-12. https://doi.org/10.5334/dsj-2021-012
14. Krause, M. (2024). *The Complete Developer: Master the Full Stack with TypeScript, React, Next. js, MongoDB, and Docker*. No Starch Press. https://books.google.com/books?hl=en&lr=&id=dnO1EAAAQBAJ&oi=fnd&pg=PR19&dq=next+js+and+mongodb&ots=H1JZxVjGPY&sig=sK0Axhkiaz4ev8akPg39NMWf2Bg
15. Kumar, A., Gawande, A., Paliwal, J., Pendse, V., Kale, S., Agarwal, A., ... & Raibagkar, S. (2025). Barriers and need for dataset sharing in the publishing of research thesis. *Iberoamerican Journal of Science Measurement and Communication*, *5*(2), 1-17. https://doi.org/10.47909/ijsmc.192
16. Metsis, S. (2020). *Cross-Sectoral Data Collaboration Platforms* (Doctoral dissertation, Tallinn University of Technology). https://digikogu.taltech.ee/et/Download/fa7cdd83-4847-4a15-b11f-5f12e0b9885a/Sektoritelesedandmekoostplatvormid.pdf
17. Mora, L. T. (2023). Development of a patient and customer management application using Dataverse Dynamics365 and Power Platforms. https://docta.ucm.es/entities/publication/19202364-e783-40fc-8703-38c8d5024198
18. Mushi, G. E., Pienaar, H., & van Deventer, M. (2020). Identifying and implementing relevant research data management services for the library at the University of Dodoma, Tanzania. *Data Science Journal*, *19*, 1-1. https://doi.org/10.5334/dsj-2020-001
19. Patel, V. (2023). Analyzing the impact of Next. JS on site performance and SEO. *International Journal of Computer Applications Technology and Research*, *12*(10), 24-27. https://www.researchgate.net/profile/Vishal-Patel-121/publication/376046436_Analyzing_the_Impact_of_NextJS_on_Site_Performance_and_SEO/links/656830bace88b870311fcabc/Analyzing-the-Impact-of-NextJS-on-Site-Performance-and-SEO.pdf
20. Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. I. (2017). Zenodo in the spotlight of traditional and new metrics. *Frontiers in research metrics and analytics*, *2*, 13. https://doi.org/10.3389/frma.2017.00013
21. Razzouk, R., & Shute, V. (2012). What is design thinking and why is it important?. *Review of educational research*, *82*(3), 330-348. https://doi.org/10.3102/0034654312457429

22. Soules, C. A., & Goodson, G. R. (2003). Metadata efficiency in versioning file systems. In the 2nd *USENIX Conference on File and Storage Technologies (FAST 03)*. https://doi.org/10.21236/ADA461077
23. Srivastava, S., Shukla, H., Landge, N., Srivastava, A., & Jindal, D. (2024). A Comprehensive Review of Next. js Technology: Advancements, Features, and Applications. *Features, and Applications (May 16, 2024)*.. https://doi.org/10.2139/ssrn.4831070
24. Toma, M., Bönisch, C., Löhnhardt, B., Kelm, M., Bohnenberger, H., Winkelmann, S., ... & Kesztyüs, T. (2024). Research collaboration data platform ensuring general data protection. *Scientific Reports*, *14*(1), 11887. https://doi.org/10.1038/s41598-024-61912-8
25. Turner, G., Culhane, M., Delwar, I., & Sizemore, J. (2017). Data Archive-Zenodo. https://vtechworks.lib.vt.edu/items/6f0eac39-b71b-4e19-92d4-e51c393f532e
26. Umbrich, J., Neumaier, S., & Polleres, A. (2015, August). Quality assessment and evolution of open data portals. In *2015 3rd international conference on future internet of things and cloud* (pp. 404-411). IEEE.. https://doi.org/10.1109/FiCloud.2015.82
27. Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., ... & Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC public health*, *14*(1), 1144. https://doi.org/10.1186/1471-2458-14-1144
28. Waidelich, L., Richter, A., Kölmel, B., & Bulander, R. (2018, June). Design thinking process model review. In *2018 IEEE international conference on engineering, technology and innovation (ICE/ITMC)* (pp. 1-9). IEEE.. https://doi.org/10.1109/ICE.2018.8436281
29. Wiehe, S. E., Rosenman, M. B., Chartash, D., Lipscomb, E. R., Nelson, T. L., Magee, L. A., ... & Aalsma, M. C. (2018). A solutions-based approach to building data-sharing partnerships. *eGEMs*, *6*(1), 20. https://doi.org/10.5334/egems.236
30. Willoughby, C., & Frey, J. G. (2022). Data management matters. *Digital Discovery*, *1*(3), 183-194.. https://doi.org/10.1039/D1DD00046B
31. Winn, J. (2013). Open data and the academy: An evaluation of CKAN for research data management. https://repository.lincoln.ac.uk/articles/conference_contribution/Open_data_and_the_academy_an_evaluation_of_CKAN_for_research_data_management/25187387/2?file=44466191
32. Yang, J. J., Li, J. Q., & Niu, Y. (2015). A hybrid solution for privacy preserving medical data sharing in the cloud environment. *Future Generation computer systems*, *43*, 74-86. https://doi.org/10.1016/j.future.2014.06.004

Appendix

Ethics Committee Approval



This is to certify that
Project No.: **OTHER2025-7-311364**
Project Title: **Design and Implementation of a Data Repository for Collaborative Research: A Pilot Platform for Dataset Management and Continuity**
Principal Researcher: **Ilyass Jannah**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 7/31/2025.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of
- Any significant change to the project and the reason for that change;
- Any unforeseen events or unexpected developments that merit notification;
- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 7/31/2025

NOVA IMS Ethics Committee
ethicscommittee@novaims.unl.pt

Interview Questions

01. What does data quality mean to you?
02. What challenges do you face regarding data quality in your work or studies?
03. How do you ensure the quality of data before performing analysis?
04. Do you use any tools or frameworks to assess and improve data quality? If so, which ones?
05. How do you handle missing, inconsistent, or duplicate data?

06. Do you think data quality impacts the outcome of data analysis or machine learning models? If yes, how?
07. Have you ever experienced a situation where poor data quality led to wrong insights or decisions? Can you describe it?
08. What suggestions do you have for improving data quality in organizations or academic research?
09. How do you document your data cleaning and preparation process?
10. In your opinion, who should be responsible for data quality in a project or organization?
11. How important do you think metadata is in relation to data quality?
12. What advice would you give to someone starting to work with data to ensure they maintain high data quality standards?
13. Do you think current tools and technologies are sufficient to ensure high data quality? Why or why not?
14. Have you been part of a project where data quality checks were automated? If yes, how was that implemented?
15. How do you think data quality considerations differ between academic research and business settings?
16. Where do you see the future of data quality management heading in the next 5–10 years?

SURVEY QUESTIONS

Part 1. Current Data Management Practices:

01. How do you currently store and manage your research datasets?
02. How would you rate your experience using your dataset storage solution?
03. How much of an issue does each of the following cause when working when organizing or retrieving datasets?
04. How do you keep track of dataset versions or updates in your research?
05. Do you collaborate with others on research datasets?
06. If yes, how do you typically share or update data with collaborators?
07. Have you ever lost or accidentally overwritten important research data?

Part 2. Impact on Research Workflow:

01. On average, how much time per week do you spend collecting data for your research?
02. What obstacles do you face when making your data reusable?
03. How often do you experience dataset duplication or conflicting versions in your team or project group?
04. How valuable would access to early versions of datasets be for your analytical work?
05. Does the data you work with have any restrictions like NDA or confidentiality issues?

Part 3. Expectations & Potential Impact of a centralized data repository:

01. Which features would be most valuable in a centralized data repository? (Choose up to 3)
02. How would a well-organized repository impact your research workflow?
03. What concerns do you have about transitioning to a new data repository system?
04. How likely are you to use a centralized research data repository if it meets your needs?

Part 4. Additional Comments: