INGENIUM
European University

Deliverable 3.5

Open Data & Open Science Repository v2

Work package 3 – Digital INGENIUM

Co-funded by
the European Union

Call: ERASMUS-EDU-2022-EUR-UNIV (EUROPEAN UNIVERSITIES)
Topic: ERASMUS-EDU-2022-EUR-UNIV-2

Proposal number: 101090042
Proposal acronym: INGENIUM

Project duration: from 1 January 2023
to 31 December 2026

COORDINATOR
University of Oviedo (UNIOVI), Spain

PARTNERS
Medical University - Sofia (MUS), Bulgaria
University of Crete (UoC), Greece
Karlsruhe University of Applied Sciences (HKA), Germany
South-Eastern Finland University of Applied Sciences (XAMK), Finland
University 'G. d'Annunzio', Chieti-Pescara (Ud'A), Italy
University of Skövde (HS), Sweden
Munster Technological University (MTU), Ireland
University of Rouen, Normandy (URN), France
'Gheorghe Asachi' Technical University of Iasi (TUIASI), Romania

Project URL: https://ingenium-university.eu/

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor the granting authority can be held responsible for them.

Table of contents

Document information

Project number	101090042	Acronym	INGENIUM
Full title	INGENIUM Alliance European Universities		

Deliverable number: 3.5	Open Data & Open Science Repository v2
Work package number: 3	Digital INGENIUM
Tasks related:	Task 3.4 Open Data and Open Science
WP lead beneficiary	University of Crete

Due date	M(42) – June 2026	
Delivery date	30/06/2026	
Status	Version: 1.1	Draft ☐ Final ☑
Type	R-Document, report ☐ DEC-Websites, patent filings, videos, etc. ☑ OTHER ☐	
Dissemination level	SEN-Sensitive ☐ PU-Public ☑	

Description of the deliverable (3-5 lines)	Open access platform associated to the Alliance portal adaptable to different devices that adhere to FAIR principles
Key words	Open Data, Open Science, Repository

Document history

Date	Version	Prepared by	Description
27/03/2026	1.0	Anastasaki Maria	First draft
02/04/2026	1.0	Tzikoulis Vasileios	Second draft
05/04/2026	1.0	Tzikoulis Vasileios	Third draft
15/06/2026	1.1	Georgios Chalkiadakis	Final Version

Definitions & Acronyms

AAI	Authentication and Authorization Infrastructure (AAI)
CKAN	Comprehensive Knowledge Archive Network
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
ORCID	Open Researcher and Contributor ID
RBAC	role-based access control
UNF	Universal Numeric Fingerprint

EXECUTIVE SUMMARY

The migration from CKAN to Dataverse represents a strategic transition from a general-purpose data cataloguing platform to a research-oriented data infrastructure aligned with the principles of Open Science and FAIR data management.
This deliverable provides a comprehensive technical and academic justification for this transition, demonstrating how Dataverse enables the treatment of datasets as first-class scholarly outputs using persistent identifiers, structured metadata, and versioning mechanisms.
Furthermore, the report situates migration within the broader European policy landscape, including the European Open Science Cloud (EOSC) and European requirements, highlighting how Dataverse supports interoperability, reproducibility, and long-term data stewardship.
The analysis concludes that the adoption of Dataverse significantly enhances the institution's capacity to manage, curate, and disseminate research data in a sustainable and standards-compliant manner.

1. Introduction

Research data has evolved into a primary scholarly asset, requiring infrastructures that support not only storage and dissemination but also reproducibility, citation, and long-term preservation.

Traditional open data platforms such as CKAN have been widely adopted for government and public data portals due to their flexibility and extensibility. However, their underlying architecture is primarily designed for data cataloguing rather than for managing the full lifecycle of research data.

As research practices increasingly demand compliance with FAIR principles [1] and alignment with European Open Science policies, limitations of such platforms become evident. These include the lack of native support for persistent identifiers, standardized metadata schemas, structured curation workflows, and dataset versioning.

In response to these challenges, INGENIUM University initiated the transition to Dataverse, a platform specifically designed for research data management.

This migration reflects a broader shift toward treating datasets as citable, reusable, and interoperable research objects, integrated within a structured and policy-compliant digital ecosystem.

2. Overview of the Dataverse Platform

The Dataverse [2] platform is an open-source research data repository system designed to support the publication, citation, sharing, and long-term preservation of research data. Developed and maintained by the Institute for Quantitative Social Science at Harvard University, Dataverse has evolved into a widely adopted solution for academic institutions, research organizations, and data infrastructures worldwide.

At its core, Dataverse is designed to treat datasets as first-class scholarly outputs. It enables researchers to deposit datasets accompanied by rich metadata, assign persistent identifiers, and make data available under controlled or open access conditions. The platform supports the full research data lifecycle, including submission, curation, publication, versioning, and reuse.

2.1 Public Availability and Access Model

Dataverse can be deployed either as a public open-access repository or as a restricted institutional platform, depending on the policies and requirements of the hosting organization. In most academic deployments, including the INGENIUM use case, the platform is configured to support open science practices while maintaining flexibility for restricted or sensitive data.

Specifically, Dataverse supports multiple access modes:

- Open access datasets, which are freely available to all users without restrictions
- Restricted datasets, where access is granted upon request or based on user roles
- Embargoed datasets, which become publicly available after a defined period

Importantly, even when data files are restricted, the associated metadata remains publicly visible. This ensures that datasets remain discoverable and citable, in accordance with FAIR principles.

Dataverse repositories are typically accessible through a dedicated institutional web interface (e.g., https://opendata.ingenium-university.eu/), providing a stable and persistent access point for both human users and machine-based services.

2.2 Licensing Model and Cost Structure

The Dataverse software is distributed as free and open-source software under the Apache 2.0 license. This means that institutions can download, install, use, and modify the platform without any licensing fees.

However, while the software itself is free, operating a Dataverse repository involves institutional costs related to infrastructure and maintenance. These typically include:

- Server infrastructure and storage capacity
- System administration and technical support
- Backup, security, and long-term preservation services
- Persistent identifier registration services, commonly provided via DataCite

In the case of INGENIUM University, the hosting, technical operation, and ongoing maintenance of the platform are provided by the University of Crete, specifically through its Digital Governance Unit and associated technical staff. This institutional support ensures the reliable operation, security, and long-term sustainability of the repository.

As such, Dataverse follows a community-driven, non-commercial model, where the software is freely available, while operational sustainability is ensured through institutional infrastructures and dedicated technical expertise.

2.3 Core Functional Capabilities

Dataverse provides a comprehensive set of functionalities that distinguish it from general-purpose data catalogues. These capabilities are specifically designed to support research data as reusable and citable outputs:

- Assignment of persistent identifiers (DOIs) for datasets
- Support for standardized metadata schemas (e.g., DataCite, DDI, Dublin Core)
- Dataset versioning and provenance tracking
- Advanced search and indexing mechanisms
- Role-based access control and curation workflows
- APIs for interoperability and machine access

These features enable the platform to function not only as a repository but as an integral component of the research ecosystem.

2.4 Global Adoption and Ecosystem

Dataverse is widely used by universities, research institutions, and data infrastructures across Europe and internationally. Its adoption is closely aligned with Open Science initiatives and infrastructures such as the European Open Science Cloud [3], which promote interoperability, data sharing, and FAIR-compliant research practices.

The platform benefits from an active global community of developers and users, contributing to continuous improvement, regular updates, and the development of new features. This community-driven ecosystem ensures that Dataverse remains a sustainable and evolving solution for research data management.

2.5 Relevance for INGENIUM University

The selection of Dataverse as the target platform reflects its strong alignment with the strategic objectives of INGENIUM University. Its open-source nature, support for FAIR principles, and compatibility with European research infrastructures make it particularly suitable for institutions aiming to establish a modern, interoperable, and policy-compliant data repository.

3.Strategic Alignment with European Policies

The migration from CKAN to Dataverse is closely aligned with key European research and data governance frameworks, particularly the European Open Science Cloud (EOSC) and the Horizon Europe Open Science policy. These initiatives establish a clear requirement for research infrastructures to support FAIR (Findable, Accessible, Interoperable, Reusable) data principles by design, rather than as an afterthought.

EOSC envisions a federated ecosystem of interoperable data repositories and services, enabling seamless access, sharing, and reuse of research data across institutional and national boundaries. Within this context, platforms are expected to provide standardized metadata, persistent identifiers, and machine-readable interfaces to facilitate integration and data exchange.

CKAN, while effective as a general-purpose data catalogue, does not natively fulfill these requirements without significant customization. In particular, the absence of built-in support for persistent identifiers, standardized academic metadata schemas, and structured curation workflows limits its ability to operate as a fully compliant EOSC node.

In contrast, Dataverse is inherently aligned with these expectations. It provides native integration with DOI services through DataCite [4], supports standardized metadata schemas such as DDI [5] and Dublin Core [6], and enables structured workflows for dataset curation and publication. These features ensure that datasets are not only accessible but also interoperable and reusable within a broader European research infrastructure.

Furthermore, Horizon Europe mandates the adoption of FAIR-by-design data management practices, requiring institutions to implement infrastructures that support reproducibility, transparency, and long-term preservation. Dataverse directly supports these requirements

through dataset versioning, provenance tracking, and rich metadata documentation, thereby enabling compliance with Data Management Plan (DMP) obligations.

Therefore, the transition to Dataverse should be understood not only as a technological upgrade but as a strategic alignment with European research policy, ensuring that INGENIUM University can effectively participate in the evolving Open Science ecosystem.

4.Architectural Comparison

CKAN architecture is catalog-centric and modular. This section provides an in-depth academic and technical analysis of the migration process, examining architectural paradigms, metadata standards, and interoperability requirements.

It emphasizes the transition from data cataloguing approaches toward integrated research data lifecycle management, where datasets are treated as scholarly outputs with persistent identifiers, versioning, and reproducibility guarantees.

Furthermore, the discussion situates the migration within the broader European Open Science ecosystem, highlighting compliance with FAIR principles and alignment with European requirements.

The analysis demonstrates how modern infrastructures must enable seamless interaction between data, researchers, and computational environments, ensuring long-term accessibility, traceability, and reuse across disciplines.

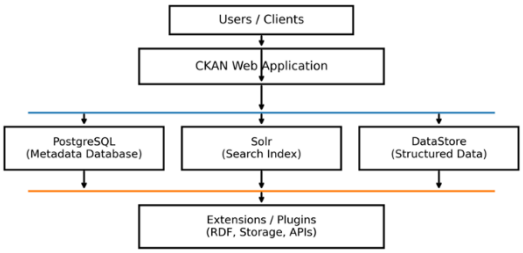


Figure 1 CKAN Architecture Diagram

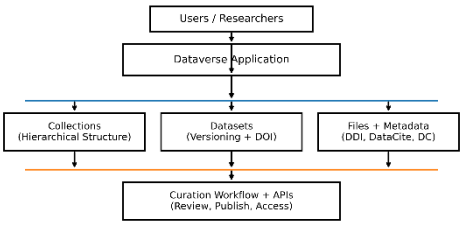


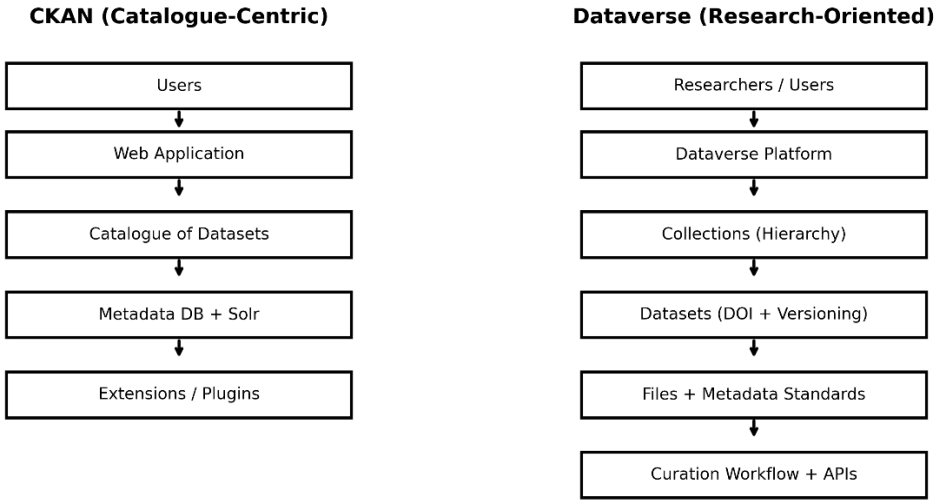
Figure 2 Dataverse Architecture Diagram

Dataverse introduces research-oriented abstractions and DOI integration. This section provides an in-depth academic and technical analysis of the migration process, examining architectural paradigms, metadata standards, and interoperability requirements.

It emphasizes the transition from data cataloguing approaches toward integrated research data lifecycle management, where datasets are treated as scholarly outputs with persistent identifiers, versioning, and reproducibility guarantees.

Furthermore, the discussion situates migration within the broader European Open Science ecosystem, highlighting compliance with FAIR principles and alignment with Horizon Europe requirements.

The analysis demonstrates how modern infrastructures must enable seamless interaction between data, researchers, and computational environments, ensuring long-term accessibility, traceability, and reuse across disciplines.



5. FAIR Principles Implementation

The implementation of FAIR principles constitutes a central requirement for modern research data infrastructures. Migration from CKAN to Dataverse enables the operationalization of these principles through built-in functionalities that support the full lifecycle of research data.

Findability is ensured in Dataverse through the systematic use of persistent identifiers, particularly Digital Object Identifiers (DOIs) assigned via DataCite. Each dataset is associated with rich, structured metadata that is indexed and searchable, enabling both human users and machine agents to discover datasets efficiently. In contrast, CKAN does not natively enforce the use of persistent identifiers, which may limit long-term discoverability and citation.

Accessibility in Dataverse is achieved through standardized access mechanisms, including web interfaces and machine-readable APIs. The platform supports configurable access controls, allowing datasets to be openly available or restricted based on ethical, legal, or institutional requirements. This ensures that data remains accessible under clearly defined conditions, in alignment with FAIR principles.

Interoperability is supported through the adoption of established metadata standards such as DDI, DataCite, and Dublin Core, as well as through the use of machine-readable formats (e.g., JSON, XML). These features enable seamless integration with external systems and facilitate data exchange across repositories. CKAN, while extensible, relies on flexible metadata structures that may result in inconsistencies and reduced semantic interoperability.

Reusability is ensured through comprehensive metadata documentation, clear licensing frameworks, and dataset versioning. Dataverse enables the tracking of dataset provenance and changes over time, allowing researchers to understand the context and evolution of the

data. Additionally, features such as the Universal Numeric Fingerprint (UNF) [7] support reproducibility by ensuring data integrity. CKAN does not provide equivalent native mechanisms for versioning and reproducibility at the dataset level.

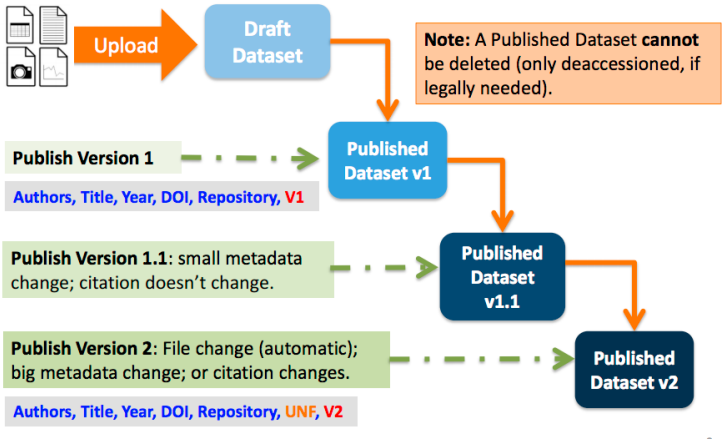


Figure 3 Dataverse Versioning

Overall, Dataverse provides a FAIR-by-design architecture, embedding these principles within its core functionality rather than requiring external extensions or custom implementations. This significantly enhances the ability of INGENIUM University to support transparent, reproducible, and reusable research.

6. Metadata Standards

Metadata standardization is a fundamental component of research data management, directly influencing the discoverability, interoperability, and reusability of datasets. The migration from CKAN to Dataverse significantly enhances the institution's ability to manage metadata in a structured, consistent, and semantically rich manner.

Dataverse provides native support for widely adopted academic metadata standards, including the Data Documentation Initiative (DDI), the DataCite Metadata Schema, and Dublin Core. Each of these standards serves a distinct role within the research data ecosystem. DDI enables detailed description of datasets, particularly in the social sciences, including variables, methodologies, and data collection processes. DataCite supports the assignment of Digital Object Identifiers (DOIs), facilitating dataset citation and integration into scholarly communication workflows. Dublin Core ensures baseline interoperability across systems by providing a common set of descriptive elements.

A key advantage of Dataverse is its ability to integrate these schemas within a unified metadata framework. This ensures that datasets are described consistently and in a machine-readable manner, enabling automated indexing, harvesting, and integration with external services. Furthermore, Dataverse supports the extension of metadata through domain-specific blocks, allowing institutions to tailor metadata to the requirements of specific disciplines such as life sciences, geospatial research, or astronomy.

In contrast, CKAN relies on a flexible metadata model based primarily on key-value pairs, which, while adaptable, may lead to inconsistencies across datasets and reduced semantic

interoperability. Although extensions and plugins can be used to introduce structured metadata, these approaches often require additional configuration and do not guarantee standardization across the platform.

The adoption of Dataverse therefore represents a shift from a flexible but loosely structured Metadata approach toward a standardized and semantically consistent metadata infrastructure. This transition is essential for enabling interoperability across repositories, supporting machine-actionable data workflows, and ensuring compliance with FAIR principles and European Open Science requirements.

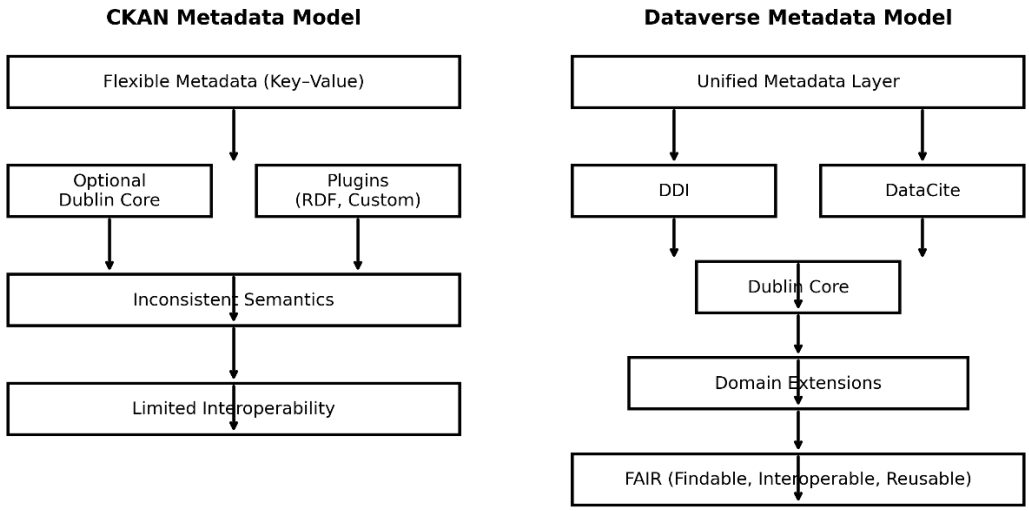


Figure 4 Metadata comparison diagram

7. Dataset Lifecycle and Curation

The management of the research data lifecycle is a critical component of modern data infrastructures, directly influencing data quality, reproducibility, and long-term usability. The migration from CKAN to Dataverse introduces a structured and policy-driven approach to dataset curation, which is essential for ensuring that research data meets academic and institutional standards.

Dataverse implements a well-defined dataset lifecycle that encompasses data submission, metadata annotation, review, publication, and versioning. Researchers (contributors) are responsible for uploading datasets and providing initial metadata descriptions, while curators perform validation and quality control prior to publication. This “submit for review” workflow ensures that datasets are not published without verification, thereby improving consistency, completeness, and reliability.

A key feature of Dataverse is the integration of metadata validation and curation into the publication process. Curators can assess metadata completeness, verify file formats, ensure compliance with ethical and legal requirements, and enforce the use of standardized

vocabularies. This structured workflow significantly enhances the quality of published datasets and reduces the risk of incomplete or inconsistent data.

In contrast, CKAN follows a more permissive publication model, where datasets can be published with minimal validation and without a formal review process. While this approach supports rapid data dissemination, it may lead to variability in metadata quality and reduced reproducibility.

Furthermore, Dataverse supports dataset versioning, allowing updates to be tracked over time while preserving previous versions. This is essential for scientific reproducibility, as it enables researchers to reference specific dataset versions used in analyses. Combined with persistent identifiers (DOIs), this ensures that datasets remain citable and traceable throughout their lifecycle.

Overall, Dataverse transforms the dataset lifecycle from a simple upload-and-publish process into a structured, curated workflow that supports high-quality, reproducible, and reusable research data, in alignment with FAIR principles and European Open Science requirements.

8. Advanced Technical Capabilities

Dataverse provides a range of advanced technical capabilities that extend beyond basic data storage and dissemination, supporting reproducibility, data integrity, and secure data sharing. These features are critical for enabling research data to function as reliable and verifiable scientific outputs.

One of the key capabilities of Dataverse is dataset versioning, which allows datasets to be updated while preserving previous versions. Each version is assigned a unique identifier, ensuring that researchers can reference and access the exact dataset used in their analyses. This capability is essential for reproducibility, as it enables the verification of scientific results over time and supports transparent research practices.

In addition, Dataverse implements the UNF, a cryptographic hash that uniquely represents the content of tabular data. The UNF enables researchers to verify that datasets have not been altered, even when stored or transferred across different systems. This provides a robust mechanism for ensuring data integrity and reproducibility, particularly in data-intensive research domains.

Dataverse also supports advanced data ingestion processes, particularly for tabular data formats such as CSV, Stata, and SPSS. During ingestion, data can be normalized and enriched with additional metadata, enabling improved searchability and analysis. This functionality enhances the usability of datasets and facilitates integration with analytical tools.

Another important capability is the integration of privacy-preserving techniques, such as those enabled through OpenDP[7]. These tools allow sensitive datasets to be shared while minimizing the risk of disclosure, supporting compliance with data protection regulations and ethical research standards.

In contrast, CKAN does not provide native support for dataset versioning, reproducibility mechanisms such as UNF, or integrated privacy-preserving tools. While some functionalities

can be introduced through extensions, these are not part of the core architecture and require additional configuration.

Overall, Dataverse offers a comprehensive set of advanced capabilities that support reproducible, secure, and high-quality research data management, reinforcing its suitability as a research-oriented data infrastructure aligned with FAIR and European Open Science requirements.

9. Identity and Access

Identity and access management constitute a critical component of research data infrastructures, particularly within the context of federated European research ecosystems. The migration from CKAN to Dataverse enables the adoption of standardized authentication and authorization mechanisms that support institutional integration, secure access, and persistent researcher identification.

Dataverse integrates with federated authentication systems such as Shibboleth [7, 8] and eduGAIN [9], allowing users to access the platform using their institutional credentials. This approach aligns with the Authentication and Authorization Infrastructure (AAI) model promoted within the European Open Science Cloud (EOSC), facilitating seamless access across institutions and reducing the need for local account management.

In addition to institutional authentication, Dataverse supports integration with ORCID (Open Researcher and Contributor ID) [10], enabling the association of datasets with persistent researcher identifiers. This ensures that research outputs are unambiguously linked to their creators, enhancing attribution, discoverability, and academic recognition. The use of ORCID also supports interoperability with external systems, including publication repositories and research information systems.

Dataverse implements a role-based access control (RBAC) model, allowing fine-grained management of permissions at the level of dataverses, datasets, and files. Roles such as administrators, curators, and contributors can be assigned specific privileges, ensuring that access and modification rights are controlled in accordance with institutional policies.

In contrast, CKAN provides more limited support for federated authentication and lacks native integration with persistent researcher identifiers such as ORCID. While authentication plugins can be implemented, these are not part of the core platform and may introduce additional complexity.

Overall, Dataverse enables a robust and standards-compliant identity and access management framework, supporting secure collaboration, proper attribution of research outputs, and alignment with European Open Science infrastructures.

10. Organizational Structure

The organizational structure of a research data repository plays a crucial role in enabling effective data management, governance, and scalability. The migration from CKAN to Dataverse introduces a hierarchical model that allows the data infrastructure to closely reflect the institutional structure of INGENIUM University.

Dataverse is built around a nested hierarchy of collections (dataverses), datasets, and files. Collections can be organized to represent faculties, departments, research groups, or projects, enabling a clear mapping between the repository structure and the organizational units of the institution. This hierarchical model facilitates intuitive navigation, ownership, and management of datasets across different levels of the organization.

A key advantage of this approach is the ability to delegate administrative responsibilities. Different units can manage their own collections, assign roles to users, and enforce local policies, while still operating within a unified institutional framework. This decentralized governance model enhances scalability and allows the repository to grow organically as new research units and projects are added.

Furthermore, the hierarchical structure supports consistent policy enforcement and metadata standardization across the institution. Templates, permissions, and workflows can be defined at higher levels and inherited by subordinate collections, ensuring alignment with institutional and regulatory requirements.

In contrast, CKAN primarily operates on a flat data catalogue model, where datasets are organized without a strong hierarchical structure. While groups and organizations can be defined, they do not provide the same level of hierarchical depth, inheritance, and governance control as Dataverse collections. This limitation can make it more difficult to manage complex institutional structures and enforce consistent data management practices.

The adoption of Dataverse therefore enables INGENIUM University to implement a structured and scalable organizational model, supporting both centralized governance and decentralized management, in alignment with best practices in research data management and European Open Science frameworks.

11. Comparative Evaluation

A comparative evaluation between CKAN and Dataverse highlights fundamental differences in their design philosophy, functional capabilities, and suitability for research data management. While both platforms are open-source and support data dissemination, their intended use cases and architectural approaches diverge significantly.

CKAN is primarily designed as a general-purpose open data catalogue, widely adopted in governmental and public sector contexts. Its strengths lie in its flexibility, extensibility, and ability to rapidly publish datasets. However, this flexibility often comes at the expense of standardization, particularly in relation to metadata, dataset versioning, and curation workflows. As a result, CKAN is well-suited for publishing heterogeneous datasets but less effective in supporting structured, reproducible research data.

In contrast, Dataverse is explicitly designed as a research data repository, where datasets are treated as citable, versioned, and curated scholarly objects. Its architecture incorporates persistent identifiers (DOIs), standardized metadata schemas, and structured workflows for data submission and review. These features enable consistent data management practices and support reproducibility and reuse.

From a metadata perspective, Dataverse provides native support for established academic standards such as DDI, DataCite, and Dublin Core, ensuring semantic interoperability across systems. CKAN, while extensible, relies on flexible metadata structures that may lead to inconsistencies and reduced interoperability.

In terms of data lifecycle management, Dataverse implements a formal curation workflow, including submission, review, and publication stages, whereas CKAN typically allows direct publication with limited validation. This distinction has a direct impact on data quality and reliability.

Furthermore, Dataverse supports dataset versioning and reproducibility mechanisms such as the Universal Numeric Fingerprint (UNF), which are not available in CKAN as core features. These capabilities are essential for ensuring scientific transparency and verification of results.

Overall, the evaluation demonstrates that CKAN remains an effective platform for open data dissemination, but it does not fully address the requirements of research data as scholarly outputs. Dataverse, by contrast, provides a comprehensive, standards-compliant, and research-oriented solution, making it significantly more suitable for academic institutions and aligned with FAIR principles and European Open Science policies.

12. Migration Implementation

The migration from CKAN to Dataverse at INGENIUM University was implemented as a structured, multi-phase process combining infrastructure deployment, system configuration, and data transformation. Rather than a simple platform replacement, this effort established a fully operational research data repository aligned with FAIR principles and European Open Science requirements.

The implementation began with the installation and deployment of Dataverse in a production-grade environment. This required the coordination of multiple system components and the establishment of a stable technical foundation capable of supporting long-term research data management. The deployment architecture was carefully designed to ensure reliability, scalability, and compliance with institutional IT policies.

At the infrastructure level, the following core components were installed and configured:

- The Dataverse application running on an enterprise-grade application server (e.g., Payara/Glassfish), ensuring high availability and performance

- A PostgreSQL relational database for the persistent storage of metadata and system configuration

- An Apache Solr indexing service to support advanced search and dataset discoverability
- Secure storage systems for dataset files, including backup and redundancy mechanisms
- HTTPS configuration and domain integration to provide secure and persistent web access

In parallel, integration with external services was established, most notably for the assignment of persistent identifiers. DOI registration workflows were configured through DataCite, enabling datasets to be uniquely identified, cited, and resolved globally.

Following the successful deployment of the core infrastructure, significant effort was dedicated to system configuration and institutional customization. These activities were essential in order to adapt the platform to the specific operational, academic, and governance requirements of INGENIUM University. The configuration process extended beyond basic setup and involved the alignment of the platform with research data management policies and FAIR compliance objectives.

Key configuration activities included:

1. Activation and customization of metadata schemas, including DataCite, DDI, and Dublin Core, to ensure semantic consistency and interoperability
2. Definition of the institutional role and permissions model, implementing a Role-Based Access Control (RBAC) framework aligned with governance structures
3. Customization of the user interface, including institutional branding, navigation structure, and multilingual support where required
4. Configuration of authentication mechanisms, including federated login (e.g., Shibboleth/eduGAIN) and integration with ORCID for researcher identification
5. Enablement of API access and interoperability features to support machine-to-machine communication and external integrations

A central and technically demanding phase of the migration involved the transfer of existing data from the CKAN platform. This process required not only the movement of files but also the transformation of metadata into a structured and standardized format compatible with Dataverse. Given the differences between the flexible metadata model of CKAN and the schema-based approach of Dataverse, particular attention was given to ensuring semantic alignment and data quality.

The data migration process included:

- Extraction of datasets and associated metadata from the CKAN repository
- Mapping and transformation of metadata fields to Dataverse-compatible schemas

- Enrichment and normalization of metadata to meet FAIR principles and improve discoverability
- Controlled ingestion of datasets into the Dataverse environment using batch processes and APIs
- Validation and quality assurance procedures to verify data integrity, metadata completeness, and functional correctness

Throughout this phase, iterative testing and validation cycles were performed in order to ensure that migrated datasets retained their scientific value and usability. This included checking file accessibility, metadata accuracy, and proper indexing within the search system.

The overall implementation required substantial institutional effort and coordination across multiple roles. The migration was not purely technical but also involved data curation expertise and organizational alignment. The main contributors to the process included:

- System administrators responsible for infrastructure deployment and maintenance
- Developers supporting configuration, customization, and automation tasks
- Data curators responsible for metadata validation and quality assurance
- Research support staff facilitating communication with data providers and ensuring compliance with policies

The migration followed a phased methodology, including planning, pilot implementation, evaluation, and full-scale deployment. This approach allowed the identification and mitigation of potential risks while ensuring a smooth transition from the legacy system to the new infrastructure.

Migration to Dataverse represents a significant technical and organizational achievement. The level of effort invested reflects the complexity of transitioning to a FAIR-compliant, research-oriented data infrastructure and demonstrates the institution's commitment to sustainable, high-quality research data management.

13. Outcomes and Impact of Migration

The transition from CKAN to Dataverse has led to a substantial transformation in the way research data is managed, curated, and disseminated within INGENIUM University. The adoption of a research-oriented repository has not only addressed the technical limitations of the previous system but has also enabled the institution to align its data management practices with contemporary academic standards and European Open Science policies.

One of the most significant outcomes of migration is the establishment of a structured and controlled dataset lifecycle. Unlike the previous environment, where datasets could be published with minimal validation, the Dataverse platform enforces a workflow that includes submission, metadata enrichment, curation, and formal publication. This ensures that

datasets meet defined quality standards before becoming publicly available, thereby improving their reliability, consistency, and long-term usability.

In practical terms, the new system introduces several key improvements in research data management:

- The systematic assignment of persistent identifiers (DOIs), enabling datasets to be formally cited and integrated into scholarly communication
- The implementation of dataset versioning, allowing changes to be tracked over time while preserving previous versions
- The integration of structured metadata schemas, ensuring semantic consistency and machine-readability
- The application of curation workflows that improve metadata quality and enforce institutional standards

These features collectively elevate datasets from simple digital assets to fully recognized scholarly outputs.

A further important impact concerns the enhancement of reproducibility and scientific transparency. Through built-in versioning mechanisms and provenance tracking, researchers are now able to reference specific dataset versions and verify how data has evolved over time. This capability is essential for validating research results and supports the broader movement toward reproducible science. In addition, mechanisms such as the Universal Numeric Fingerprint (UNF) contribute to ensuring data integrity by allowing verification that datasets have not been altered.

From a policy perspective, migration has enabled the effective operationalization of FAIR principles. The platform supports improved discoverability through indexed and richly described metadata, ensures accessibility through clearly defined access conditions, enhances interoperability through the use of standardized schemas, and promotes reusability through licensing, documentation, and version control. These capabilities position INGENIUM University to actively participate in European research infrastructures such as the European Open Science Cloud.

An additional important outcome of this transition is the alignment and convergence with similar institutional initiatives at the national level. In particular, the migration contributes to harmonization with the repository infrastructure of University of Crete, which has also transitioned from CKAN to Dataverse. This convergence facilitates interoperability, knowledge exchange, and the adoption of common standards and practices across institutions. By operating on a shared technological and conceptual framework, institutions are better positioned to collaborate, exchange data, and integrate their repositories within broader national and European research ecosystems.

Migration has also had a measurable impact on the visibility and dissemination of research outputs. By assigning persistent identifiers and supporting integration with researcher identification systems such as ORCID, datasets become more easily discoverable and

attributable. This increases their citation potential and facilitates their inclusion in global discovery services, thereby extending the reach of institutional research beyond traditional publication channels.

In addition to research-related benefits, the institution has gained important organizational and operational advantages. The adoption of a structured repository has improved governance over research data, enabling clearer allocation of responsibilities and more consistent enforcement of policies. The hierarchical organization of collections allows different departments and research groups to manage their data independently while remaining within a unified institutional framework.

Key institutional benefits include:

- Improved control and oversight of research data assets
- Enhanced data quality through standardized curation processes
- Greater scalability to support growing volumes of research data
- Support for interdisciplinary collaboration through shared infrastructure

The outcomes of the migration extend beyond technical improvements, contributing to a broader cultural shift toward open, transparent, and reproducible research practices. The adoption of Dataverse has positioned INGENIUM University as a modern, standards-compliant data provider, capable of supporting both current research needs and future developments within the European Open Science ecosystem.

CONCLUSION

The migration from CKAN to Dataverse represents a necessary and strategically justified evolution of the INGENIUM University research data infrastructure. This transition reflects a shift from a general-purpose data cataloguing approach toward a research-oriented model that treats datasets as structured, citable, and reproducible scholarly outputs.

The analysis presented in this deliverable demonstrates that CKAN, while effective for open data dissemination, does not fully support the requirements of modern research data management. In particular, limitations related to metadata standardization, dataset versioning, curation workflows, and interoperability reduce its suitability for academic environments.

In contrast, Dataverse provides a comprehensive and integrated solution that addresses these challenges through its support for standardized metadata schemas, persistent identifiers (DOIs), structured curation workflows, and advanced reproducibility mechanisms. These features enable the implementation of FAIR principles by design and support compliance with European Open Science policies, including the European Open Science Cloud (EOSC) and Horizon Europe requirements.

Furthermore, the adoption of Dataverse enhances the institution's capacity to manage research data in a scalable, secure, and sustainable manner. The platform's hierarchical structure, federated authentication capabilities, and support for domain-specific metadata enable alignment with institutional needs while facilitating integration within broader research infrastructures.

Overall, the transition to Dataverse positions INGENIUM University as a modern, standards-compliant, and research-oriented data provider, capable of supporting open science practices, fostering data reuse, and contributing effectively to the European research ecosystem.

References

1. FAIR. FAIR Principles. Available from: https://www.go-fair.org/fair-principles/.
2. Harvard, U. Dataverse Project. Available from: https://dataverse.org/.
3. European, C. European Open Science Cloud (EOSC). Available from: https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science/european-open-science-cloud-eosc_en
4. Datacite. Available from: https://datacite.org.
5. DDIAlliance. Enhancing Data Reuse, Supporting Standards Compliance, Improving Interoperability. Available from: https://ddialliance.org/.
6. DublinCore. Available from: https://www.dublincore.org.
7. Harvard, U. Universal Numerical Fingerprint (UNF). Available from: https://guides.dataverse.org/en/latest/developers/unf/index.html.
8. Shibboleth, C.; Available from: https://www.shibboleth.net.
9. eduGAIN. Available from: https://edugain.org/.
10. ORCID. Available from: https://orcid.org.

Annex A. Roles and Permissions Framework

The Dataverse platform implements a sophisticated Role-Based Access Control (RBAC) model, providing a granular governance structure essential for an academic environment like ingenium-university.eu. Unlike more generalized catalog systems, Dataverse allows for the delegation of authority across hierarchical collections, ensuring that research data management aligns with departmental or project-specific policies.

Table A.1: Primary Functional Roles and Access Levels

Role	Level	Core Permissions & Academic Function
Administrator	Collection / Dataset	Full administrative control, including the ability to manage permissions, configure templates, and enforce institutional policies.
Curator	Collection / Dataset	Empowered to review datasets, validate metadata accuracy, and authorize the formal publication of research.
Contributor	Collection	Primarily for researchers, allows for the creation of datasets and the editing of metadata, but requires a "Submit for Review" step prior to publication.
Dataset Creator	Collection	A specialized role assigned at the collection level specifically to grant users the right to initiate new dataset entries.

File Downloader	File / Dataset	Provides read-only access to published datasets and files. Can be assigned to specific groups, such as “authenticated users".

Governance and Scalability Features

Permission Inheritance: Permissions can be automatically inherited from parent collections, reducing administrative overhead for large university structures.

Group-Based Governance: Access can be granted to specific IP groups or institutional login groups (e.g., via Shibboleth/eduGAIN), facilitating seamless access for the entire university community.

Advanced Curation Workflow: The separation of the Contributor and Curator roles ensures a peer-review-like process for data publication, a feature that CKAN typically handles through simpler, less granular extensions.

Annex B: Scientific File Formats and Tabular Ingest

Dataverse provides advanced handling for scientific data that goes beyond simple file storage. A central pillar of its superiority for ingenium-university.eu is the Tabular Data Ingest process, which converts proprietary formats into preservation-standard representations.

Table B.1: Supported Academic and Scientific Formats

Category	Supported Formats	Advanced Functionality (Dataverse Native)
Statistical Data	Stata, SPSS, R, Excel (xlsx), CSV/TSV	Automatic conversion to tab-delimited formats, variable-level metadata extraction, and Universal Numerical Fingerprint (UNF) generation.
Geospatial	GeoJSON, GeoTIFF, Shapefile, NetCDF, HDF5	Native map previews and extraction of geospatial bounding box metadata (longitude/latitude).
Astronomy	FITS (Flexible Image Transport System)	Automatic header metadata extraction aggregated at the dataset level for enhanced discoverability.
Research Code	R, Python, Jupyter, MATLAB, Stata	File-level metadata editing and integrated previewers for code scripts.
Documentation	PDF, Markdown, Text, HTML	Native browser-based previewing to allow quick evaluation of research context without downloading.

Technical Superiority in Data Ingest

Standardized Internal Representation: During ingest, Dataverse extracts variable names and labels from statistical files (SPSS, Stata, R), making every variable searchable through the Advanced Search interface.

Data Integrity: The application of UNF ensures that any change in the underlying data values is detectable, providing a guarantee of data fixity crucial for reproducibility.

Multimedia and Archive Support: The system provides built-in previewers for audio, video, and compressed (zip) files, allowing users to inspect contents before agreeing to terms of use.

Comparison with CKAN: While CKAN treats files as generic "resources," Dataverse's deep integration with statistical and domain-specific formats provides the variable-level granularity required for modern scientific inquiry.

Annex C: API Specifications and System Interoperability

The Dataverse platform architecture at ingenium-university.eu is built upon a robust, RESTful API framework. These Application Programming Interfaces (APIs) are critical for enabling seamless system integration, the automation of complex research workflows, and machine-to-machine interactions. By providing programmatic access to core repository functions, the platform ensures that datasets are not merely stored but are active components of a broader, interoperable research ecosystem.

Table C.1: Core Dataverse API Ecosystem

API Endpoint	Functional Scope	Research & Technical Utility
Native API	Repository Management	Enables programmatic dataset creation, metadata updates, file uploads, and advanced publication workflows.
Search API	Data Discovery	Allows external systems to query datasets and files using structured parameters, supporting indexing by global aggregators.
Data Access API	Resource Retrieval	Provides controlled, authenticated access to files, respecting both open and restricted access configurations.
Metrics API	Analytical Reporting	Aggregates usage data, such as download counts and views, following "Make Data Count" standards.
SWORD API	Standardized Deposit	Supports the AtomPub-based protocol for standardized, cross-platform dataset deposits.

Security and Authentication Protocols

Security within the API framework is managed using API Tokens. These tokens function as unique, sensitive credentials that allow external applications to act on behalf of a user.

Credential Integrity: Users must treat API tokens with the same level of security as passwords, as they grant permissions to add, modify, or delete data.

Token Lifecycle: The platform allows for the creation, rotation, and recreation of tokens to maintain a secure interaction environment.

Auditability: Every action performed via the API is tracked, ensuring accountability and compliance with institutional data governance policies.

Interoperability and Machine-Actionability

The Dataverse API framework prioritizes interoperability using standard web protocols (HTTP) and universal data formats, primarily JSON.

Standardized Metadata Exports: Published dataset metadata can be harvested and exported in multiple formats via API, including DataCite 4.5, DDI, Dublin Core, and Schema.org JSON-LD, facilitating discoverability in platforms like Google Dataset Search.

Signposting: The platform supports programmatic metadata retrieval via Signposting, enabling machines to navigate the relationship between datasets and their various metadata representations.

Integration with External Tools: The APIs allow for the development and connection of external "Data Exploration" and "Visualization" tools, enhancing the analytical capabilities available to the university's researchers.

Architectural Comparison: Dataverse vs. CKAN

While CKAN provides an API for general data management and cataloging, its focus remains largely on resource retrieval and portal management. In contrast, the Dataverse API framework is purpose-built for the scholarly research lifecycle, providing deep, native support for:

Structured Curation Workflows: Automating the "Submit for Review" process and curation status tracking.

Sophisticated Versioning: Programmatically tracking major and minor version changes to metadata and files.

Domain-Specific Metadata Handling: API-driven management of specialized metadata blocks (e.g., Life Sciences, Astronomy, Social Sciences).

This framework ensures that ingenium-university.eu maintains a high degree of flexibility and scalability, supporting automated data pipelines that align with the requirements of the European Open Science Cloud (EOSC) and modern FAIR data mandates.

Annex D. Matrix to be used by partners to list and to monitor the progress of their key institutional priorities related to the deliverable.

The numeration of this annex will be correlative with the rest and will always be the last one.

Institutional transformation objectives	• *Strengthening Open Science and Research Data Management* • *Enhancing Digital Transformation and Research Infrastructure* • *Building Institutional Capacity and Skills in Research Data Management* • *Increasing Visibility, Accessibility, and Impact of Research Outputs*
Barriers faced to achieve those objectives at the institutional level	• ***Legacy Systems and Technical Migration Challenges*** *A significant barrier was the transition from existing repository infrastructures, particularly the migration from CKAN to Dataverse. This process required data mapping, metadata harmonization, system reconfiguration, and ensuring the integrity and long-term preservation of existing datasets. Differences in data models and metadata standards between platforms increased the complexity of the migration process.* • ***Limited Institutional Capacity and Expertise*** *The successful deployment and operation of a modern research data repository require specialized expertise in research data management, metadata standards, repository administration, and Open Science practices. Building and maintaining these competencies within the institution required additional training and capacity-building efforts.* • ***Organizational Change and User Adoption*** *Researchers and administrative staff were already familiar with existing workflows and systems. Encouraging the adoption of a new platform required awareness-raising activities, user training, and the development of new operational procedures. Resistance to change and varying levels of digital maturity represented challenges during the implementation phase.*

	• ***Harmonization Across Multiple Institutions*** *The development of a shared repository framework within the INGENIUM alliance required coordination among institutions with different technological environments, policies, governance structures, and levels of experience in research data management. Achieving common standards and interoperable practices required substantial collaboration and consensus-building.* • ***Sustainability and Resource Constraints*** *Ensuring the long-term sustainability of the repository infrastructure required dedicated technical resources, ongoing maintenance, storage capacity, and institutional commitment. Securing sufficient human and financial resources for continuous operation and future development remained an important challenge.*
Potential Actions to be taken at the institutional level	• ***Consolidate and Expand the Dataverse Research Data Repository*** *Continue the development and institutional adoption of the Dataverse repository as the primary platform for research data management and sharing. This includes the migration of additional datasets, the creation of disciplinary collections, and the enhancement of repository services to support researchers throughout the data lifecycle.* • ***Strengthen Open Science and Research Data Management Policies*** *Review and update institutional policies related to research data management, data sharing, and Open Science in order to align with FAIR principles, European Open Science requirements, and emerging best practices. Promote the adoption of data management plans and open research workflows.* • ***Provide Training and Capacity Building*** *Organize training workshops, seminars, and guidance sessions for researchers, doctoral candidates, and administrative staff on topics such as research data management, metadata creation, repository*

	use, FAIR data principles, and Open Science practices. • ***Enhance Technical Infrastructure and Interoperability*** *Further develop the technical infrastructure supporting the repository by improving interoperability with institutional systems, persistent identifier services (e.g., DOI and ORCID), research information systems, and other national and European Open Science infrastructures.* • ***Establish a Research Data Support Network*** *Create a coordinated support structure involving the library, IT services, and research support offices to provide guidance on data management, repository usage, data publication, and compliance with funder requirements.* • ***Promote Repository Adoption and Research Visibility*** *Encourage researchers and research groups to deposit datasets and related research outputs in the institutional repository. Increase awareness of the benefits of data sharing, reproducibility, and open access to enhance the visibility and impact of institutional research.* • ***Strengthen Collaboration within the INGENIUM Alliance*** *Continue sharing expertise, best practices, and technical developments with partner universities in the INGENIUM alliance. Contribute to the harmonization of repository services, metadata standards, and Open Science practices across participating institutions.* • ***Monitor Impact and Continuous Improvement*** *Establish mechanisms for monitoring repository usage, dataset publication, user engagement, and service quality. Use the collected evidence to guide future improvements and ensure the long-term sustainability of the infrastructure and associated services.*

Actions to be taken at other levels	• *Collaborating with National Open Science and Research Infrastructure Initiatives* • *Promote Alignment with European Open Science Policies* • *Foster Collaboration with Other Universities and Research Organizations* • *Support Regional and National Capacity Building*
Responsible(s) within the instituion	***Main Responsible:*** • *University of Crete Information and Communication Technologies (ICT) Services.* • *Institutional Open Science and Research Data Management Coordination Team.* ***Other Involved Actors:*** • *University Library and Information Centre.* • *Research Support Office / Research Committee.* • *Academic departments and research groups.* • *Data stewards and repository administrators.* • *Legal and policy support unit*
Expected timeline and key milestones	***Milestone:*** *Dataverse repository fully operational and available to researchers.* *Year 1* • *Establish institutional governance and operational procedures.* • *Launch the Dataverse-based institutional repository.* • *Conduct initial user training and awareness activities.* ***Milestone:*** *Significant growth in deposited datasets and active users.* *Year 2* • *Expand repository adoption across departments and research units.* • *Integrate persistent identifiers (DOI, ORCID) and improve interoperability.* • *Deliver advanced training programmes and support services.* ***Milestone:*** *Sustainable institutional research data infrastructure integrated into the broader Open Science ecosystem.* *Year 3* • *Consolidate institutional Open Science services.* • *Evaluate repository impact and user engagement.*

	• *Share best practices and implementation outcomes at national and European level.* • *Develop a long-term sustainability and maintenance plan.*