



MDMC

Master in Data Management
and Curation

MASTER IN DATA MANAGEMENT AND CURATION

Innovation Intelligence FVG – A FAIRification Process

Supervisor:

DR. FEDERICA BAZZOCCHI,

Candidate:

ENRICO LONGATO

2025–2026





Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA

This course has been funded by the European Union – NextGenerationEU within the projects:

- *NFFA-DI cod. IR00000015 - Missione 4, “Istruzione e Ricerca” – Componente 2, “Dalla ricerca all’impresa” – Linea di investimento 3.1, “Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione” – Azione 3.1.1, “Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti” (CUP B53C22004310006)*

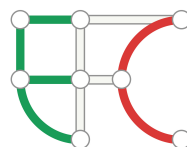
[NFFA-DI : Nano Foundries Fine Analysis - Digital Infrastructure](#)

- *EFC cod. SSU2024 – 00002 - Missione 4, “Istruzione e Ricerca” – Componente 1, “Potenziamento dell’offerta dei servizi all’istruzione: dagli asili nido all’ università” – Linea di investimento 3.4, “Didattica e competenze universitarie avanzate” – Sub-Investimento, “Rafforzamento delle scuole universitarie superiori” (CUP G97G24000100007)*

[EFC : Educating Future Citizen](#)



nffa-di



Author's Declaration

I, Enrico Longato, declare that this thesis entitled, 'Innovation Intelligence FVG – A FAIRification Process' and the work presented therein are my own.

I certify that:

- This work was performed wholly or principally during MDMC internship at the Research Institute for Technological Innovation of Area Science Park.
- If any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have relied on the published work of others, this is always clearly attributed.
- Where the thesis is based on work I have done in collaboration with others, I have made clear exactly what was done by others and what I contributed.
- With respect to AI technologies:
 - I acknowledge the use of ChatGPT - <https://chatgpt.com/> to provide information for background research and to assist in the drafting process with the creation of an outline structure for this essay.
 - I acknowledge the use of ChatGPT - <https://chatgpt.com/> to identify improvements in the writing style.
 - I acknowledge the use of ChatGPT - <https://chatgpt.com/> and Claude - <https://claude.ai>
 - I acknowledge the use of ChatGPT - <https://chatgpt.com/> and Claude - <https://claude.ai> as support tools during software development activities, including coding, debugging, and documentation assistance.
 - No content generated by AI technologies was presented as my own work. as a source of information to create materials that will be used in my own words.
- Where I have quoted from the work of others, the source is always cited. Except for such quotations, this thesis is entirely my own work.
- I have acknowledged all major sources of help.

“If the path doesn’t exist, create it.”

Robert Baden-Powell

Acknowledgments

Just a few months ago, if someone had told me that I would have reached this point, I probably would not have believed it. For this reason, I would like to deeply thank Dr. Mariarita de Luca for encouraging me to undertake this challenging and unconventional path, and Dr. Federica Bazzocchi for transmitting enthusiasm and passion when discussing my work and for trusting and guiding me throughout this thesis with long, stimulating, and insightful exchanges. Today, I am happy to collaborate with them as colleagues.

My greatest thanks go to Carolina, for always understanding and supporting me so deeply in exploring my potential and for giving me a place that I can call home and family, wherever it may be, as long as it is with her and Giacomo. Without even realizing it, Giacomo gave me both the time and the motivation needed to complete this important step, just by being there.

I would also like to thank Leyla Vesnic, with whom I worked closely in the past, and I really hope to work again in the future. Part of the work carried out together during that period later became the basis for this thesis.

Lastly, I would also like to mention my travel companions, especially Ana, Jacopo, and Smriti, with whom I shared doubts, laughs, and many moments of “I am lost”, but in the end we made it. I also have to confess that I was, and still am, genuinely impressed by how capable you all are.

Abstract

Innovation Intelligence FVG is a project developed by Area Science Park that integrates heterogeneous datasets, differing in both type and source, in order to describe the innovation capabilities of companies operating in the Friuli Venezia Giulia region. The system consists of data processing pipelines, an SQL database, and a dashboard interface. Despite the overall quality and value of the platform, the current infrastructure presents several limitations related to system complexity, fragmentation of data processing workflows, and lack of structured documentation. These issues represent the main challenge addressed in this work, which aims to improve the system while preserving the know-how and operational logic developed over time. This thesis presents the development of a FAIR-oriented proof of concept (PoC) designed to address these limitations through the application of the FAIR principles — Findability, Accessibility, Interoperability, and Reusability. The proposed PoC includes modular Python-based data pipelines, automated DCAT-compliant metadata generation, provenance tracking mechanisms, semantic crosswalks based on external ontologies, and a lightweight web architecture developed using Django, SQLite, and Docker containerization. The resulting mockup demonstrates how a FAIR-oriented framework can improve data management processes by enhancing organization, interoperability, transparency, reproducibility, and long-term maintainability, while preserving compatibility with the existing infrastructure. Although the implementation does not achieve full FAIR compliance, the work highlights both the potential and the practical challenges of FAIRification processes in complex data systems.

Contents

1	Introduction	8
1.1	Context and Motivation	8
1.2	Objectives	9
1.3	Structure of the Thesis	9
2	Innovation Intelligence FVG	11
2.1	Overview of the Project	11
2.2	System Functionality and Tools	13
3	Challenges and Limitations	15
3.1	System Complexity	15
3.2	Data Processing and Documentation Issues	16
3.3	Usability and Accessibility Limitations	16
4	FAIR Data as a Solution	18
4.1	Introduction to FAIR Principles	18
4.2	FAIR in the Context of I2FVG	19
4.3	Benefits of FAIR Data	20
5	Methodology and Implementation	21
5.1	FAIRification Approach	21
5.2	Data Processing and Cleaning	23
5.3	Documentation and Metadata	26
5.4	System Simplification and Design	30
6	Results and Discussion	33
6.1	Improvements Achieved	33
6.2	Critical Evaluation of the FAIR-oriented Approach	35

7	Conclusions and Future Developments	37
7.1	Conclusions	37
7.2	Future Developments	38

Chapter 1

Introduction

1.1 Context and Motivation

Economic and innovation-related data play an increasingly important role in understanding the evolution of a territory. Public institutions, research centers, and policymakers rely on company-related information to monitor industrial dynamics, identify innovation trends, and support strategic decision-making processes.

However, this information is rarely available in a unified form. Data describing companies, economic performance, patents, certifications, public funding, or demographic characteristics are often distributed across multiple heterogeneous sources, developed independently and following different standards, formats, and access conditions. As a consequence, transforming these fragmented datasets into a coherent and reusable knowledge infrastructure becomes a complex challenge.

Innovation Intelligence FVG (I2FVG)¹, a platform developed by Area Science Park to analyze companies operating in the Friuli Venezia Giulia region, represents a significant example of this complexity. The system integrates multiple public and proprietary data sources in order to assess structural, economic, and innovation-related characteristics of regional companies, supporting territorial analysis and policy-oriented activities.

Over time, however, the increasing complexity of the infrastructure, together with the heterogeneity of data sources and workflows, highlighted the need for a more structured and interoperable approach to data management. In this context, the FAIR principles [1] — aimed at making data Findable, Accessible, Interoperable, and Reusable — provide a useful conceptual and methodological framework for redesigning and reorganizing complex data

¹<https://www.areasciencepark.it/project/innovation-intelligence-fvg/>

ecosystems.

Within this perspective, this thesis explores how a FAIR-oriented approach can support the evolution of an existing non-FAIR infrastructure through the development of a proof-of-concept (PoC) FAIRification framework for I2FVG.

1.2 Objectives

This work originates from the need to bring order and structure to a multi-layered project that has increased in complexity over the years.

It aims to demonstrate how database systems and data-driven tools that were not originally designed according to FAIR principles can be reorganized and redesigned to become FAIR-compliant. In this context, FAIR principles represent a solution for improving coherence across the system and enabling interoperability through the adoption of standardized processes.

Using Innovation Intelligence FVG as a case study, this thesis explores the steps required to develop a FAIR-oriented PoC aimed at validating the applicability of FAIR principles to the I2FVG infrastructure. In particular, the existing data preparation processes, data ingestion pipelines, and database structure are analyzed and subsequently redesigned following FAIR principles.

This involves standardizing data processing workflows using the Python programming language, fully documenting all scripts, creating and linking metadata to existing ontologies, and developing a web application based on the Django framework supported by a SQLite database.

On top of these processes, the ultimate goal is to demonstrate the accessibility, interoperability, and reusability of the system through a GitHub repository, where the solution is containerized and provided with example (synthetic) data. This approach contributes to improving the maintainability and long-term sustainability of complex data systems.

1.3 Structure of the Thesis

This thesis is organized as follows. Chapter 2 presents an overview of the instrument Innovation Intelligence FVG, including data sources and main functionalities. Chapter 3 discusses the main challenges and limitations of the current system. Chapter 4 introduces the FAIR principles and their relevance to the project. Chapter 5 describes the methodology and implementation of the FAIRization process. Chapter 6 presents the results achieved, and

Chapter 7 concludes the thesis with a summary of the work and possible future developments.

Chapter 2

Innovation Intelligence FVG

2.1 Overview of the Project

Innovation Intelligence FVG represents a valuable example of regional data integration, where multiple data sources are combined to generate indicators related to economic performance, innovation, and sustainability. Such systems are increasingly relevant for research institutions, public administrations, and policymakers, particularly in the context of industrial policies, making I2FVG a strategic digital asset.

Innovation Intelligence FVG is a project developed by Area Science Park for the creation of a database that collects, filters, and processes publicly available information on companies operating in the Friuli Venezia Giulia region, with a particular focus on their propensity for innovation. The platform has been online since 2019 and has been fully operational since 2020.

The project originates from a need expressed by Area Science Park and shared with numerous regional stakeholders: to improve the planning, monitoring, and evaluation of activities supporting innovation within an integrated regional innovation system. The aim is to make all relevant public information easily accessible, even to non-expert users, in order to facilitate the segmentation of regional companies.

The collected information includes general data (e.g., company registry data, size, number of employees, location), sector-related and economic indicators (e.g., ATECO codes, participation in smart specialization areas, trends in turnover, labor costs and profits over time, ratings, financial indicators, export activities), as well as indicators of innovation (e.g., innovative start-ups, innovative SMEs, patents filed, regional and European funding for research, development and innovation, network contracts, and quality and environmental certifications).

A Business Analytics tool supports the visualization of data through interactive dashboards, enabling the geolocation of companies and the extraction of customized company lists based on user-defined filtering criteria.

Type of Information	Source	License
Company Certifications	Accredia ¹	Paid
Legality Rating	AGCM (Italian Competition Authority) ²	Open
EU-funded Projects (H2020 and Horizon Europe)	European Union Open Data Portal ³	Open
Company Registry Data	InfoCamere ⁴	Paid
Company Financial Data	InfoCamere ⁴	Paid
Credit Ratings	Modelfinance ⁵	Paid
Regional Innovation Funding Programmes	Regione Friuli Venezia Giulia ⁶	Open
Patents (IT and EU)	UIBM and European Patent Office ⁷	Open

Table 2.1: Data sources used by Innovation Intelligence FVG

All these data sources are processed through different types of data cleaning scripts in order to meet specific data quality standards defined by the data ingestion process within the MySQL database. The data ingestion pipeline is a complex procedure that feeds the SQL database and enables the integration of heterogeneous datasets.

During this process, several indicators are also computed based on the available data, enriching the system with derived variables that support analysis and decision-making.

The indicators included in the I2FVG system can be grouped into three main categories, as summarized in Table 2.2.

¹<https://www.accredia.it>

²<https://www.agcm.it>

³<https://data.europa.eu>

⁴<https://www.infocamere.it>

⁵<https://www.modelfinance.com>

⁶<https://www.regione.fvg.it>

⁷<https://uibm.mise.gov.it>, <https://www.epo.org>

Indicator Category	Source
Structural indicators	InfoCamere
Economic and financial indicators	InfoCamere
Innovation and strategic indicators	Multiple sources

Table 2.2: Main categories of indicators integrated in the I2FVG system

2.2 System Functionality and Tools

The system is composed of two main components: a MySQL database and a web application based on a Power BI dashboard. The database manages the integration, storage, and processing of heterogeneous data sources, while the web interface enables users to explore and visualize the information through interactive dashboards.

A testing environment is also available, mirroring the production system. This allows for data uploading and the validation of functionalities without affecting the operational version of the platform.

The overall structure of the database is illustrated in Figure 2.1.

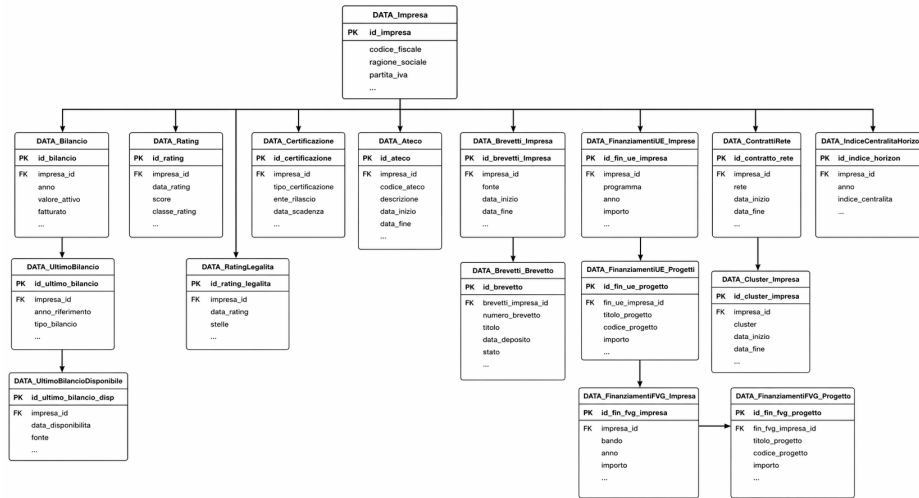


Figure 2.1: Simplified database schema of the I2FVG system

The web application is hosted at <https://i2fvg.areasciencepark.it/> and can be accessed through authentication using a username and password provided by the IT team of Area Science Park. Access to the platform is restricted to authorized personnel, including staff working at Area Science Park and public administrators of the Friuli Venezia Giulia Region.

The dashboard allows users to explore the database and apply filters based on the attributes available in the underlying data tables. The user interface is organized into thematic sections, each focusing on a specific dimension of analysis. These include: localization, sector of activity, ATECO classification¹, company structure, financial performance, financial position, credit rating, innovation and internationalization, network contracts, certifications, EU and regional (FVG) funded projects, and patents. This thematic organization enhances usability and supports targeted analysis of companies from multiple perspectives.

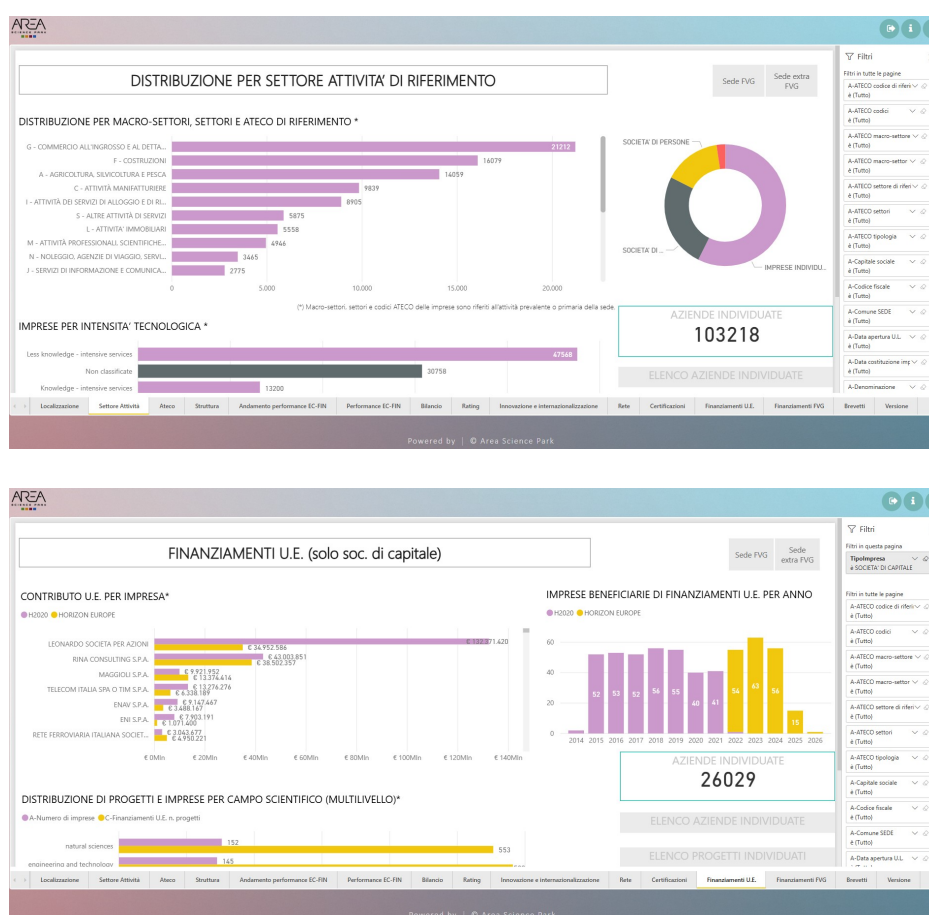


Figure 2.2: Examples of the I2FVG dashboard interface (sector of activity and EU Projects)

This overview provides the necessary context to understand the limitations of the current system, which are discussed in the following chapter.

¹<https://www.istat.it/it/archivio/17888>

Chapter 3

Challenges and Limitations

3.1 System Complexity

The innovative nature of the system lies in its development within a public research context and in its integration of both open and proprietary data sources.

The system operates in a hybrid data environment, combining openly available company data with licensed datasets. On the one hand, open data—aligned with open science principles and defined as publicly available, non-confidential, and non-personal data produced with public resources—enable transparency and support public administrations in designing evidence-based policies [2, 3]. On the other hand, licensed data enhance the overall value of the system by enriching and complementing the available information.

This combination introduces a first layer of complexity, arising from differences in data access conditions, as well as from the heterogeneity of formats, naming conventions, and standards across data sources [3].

An additional layer of complexity is related to the rigidity of the system architecture, which has required continuous modifications and ad hoc adjustments over time. Instead of evolving through a structured and maintainable design, the system has been progressively adapted through incremental fixes, leading to a fragmented and difficult-to-manage structure.

Moreover, the lack of comprehensive and consistent documentation has made knowledge transfer particularly challenging. As a result, each transition between different developers or teams has involved a partial loss of information, making it difficult to fully reconstruct the system and increasing the risk of overlooking important components.

The data processing workflows further contribute to the overall complexity of the system. Data are collected, cleaned, and transformed through

multiple scripts and procedures, often developed at different stages of the project and not consistently documented. In addition, the data cleaning pipeline is highly fragmented and implemented using multiple programming languages (mainly Python and R), while the data ingestion process relies on C++, a choice driven by the constraints of a Windows-based environment.

This dependency on a Windows environment further limits portability and flexibility, making it more difficult to adapt the system to different infrastructures or integrate it with alternative technological solutions, which is a key requirement in modern data-driven public sector systems [4].

3.2 Data Processing and Documentation Issues

Data processing operations represent a central component of the platform. As mentioned above, once data are collected, they are processed through a set of scripts that prepare them for ingestion into the database.

At present, this process presents two main issues. First, the data processing pipeline is highly fragmented, both in terms of programming languages and in the accumulation of incremental adjustments introduced over time to accommodate the ingestion phase. Second, the ingestion process itself is rigid, difficult to modify, and strongly dependent on the original developers, as well as on a specific technical environment.

Another critical issue concerns the lack of proper documentation of these processes. As a result, understanding and maintaining the system becomes time-consuming and highly prone to errors. This leads to a constant need to manually review data cleaning and ingestion steps, increasing the risk of inconsistencies and reducing overall efficiency.

These limitations undermine the reproducibility and transparency of I2FVG, which are key requirements for effective and reliable data management.

3.3 Usability and Accessibility Limitations

The usability of the platform represents a further critical layer of Innovation Intelligence FVG, as it determines the ability of users to interact with the data and fully understand its potential.

At present, despite the richness of the data available in the database, the dashboard offers a large number of visualizations, information layers, and filtering options. This abundance of functionalities can make navigation complex, causing users to feel disoriented when interacting with the system.

As a result, the platform may be perceived as overly technical and therefore difficult to use, particularly for non-expert users.

Moreover, the presence of data under restricted licenses limits accessibility for external users, reducing the potential for broader data exploration and reuse.

These limitations affect both the usability and accessibility of the system, ultimately reducing its effectiveness and limiting its potential impact.

Chapter 4

FAIR Data as a Solution

4.1 Introduction to FAIR Principles

The FAIR principles, introduced by Wilkinson et al. [1], provide a set of guidelines for managing data and making them Findable, Accessible, Interoperable, and Reusable, from which the acronym FAIR derives. This framework is now widely adopted in scientific and data-driven contexts [5], and is particularly relevant for the context of I2FVG, where heterogeneous data sources and complex processing workflows require structured data management approaches.

More specifically, Findability refers to the ability to locate data and metadata through the use of standardized and persistent identifiers, enabling efficient data discovery.

Accessibility concerns the use of clear and standardized protocols to retrieve data, which may include authentication and authorization mechanisms. It is therefore important to note that accessibility does not necessarily imply openness.

Interoperability, which is particularly relevant in the context of this work, refers to the adoption of shared standards, formats, and vocabularies that allow data from different sources to be integrated and used together. In this sense, data are made capable of "speaking" a common language.

Finally, Reusability emphasizes the importance of proper documentation, provenance information, and licensing conditions, enabling data to be reused in different contexts or to support the replication of analyses.

In practice, the FAIR principles aim to make data usable in the long term and across different contexts, enhancing their potential for integration and reproducibility [1, 5].

However, the implementation of FAIR principles in real-world systems

is often complex and challenging. Several studies have highlighted the gap between the theoretical definition of FAIR principles and their practical adoption, particularly in data-intensive environments [6, 7, 8, 9]. These works are intended as representative examples of a broader body of literature and are not exhaustive.

A large portion of this literature is concentrated in the biomedical and health data domain, where FAIR principles have been widely applied to improve data sharing, interoperability, and reproducibility [10, 11, 12]. These studies consistently report challenges related to data heterogeneity, lack of standardized metadata, and the integration of legacy systems. Again, the cited works illustrate common trends rather than providing a comprehensive overview.

In contrast, applications of FAIR principles in business and economic data systems remain relatively limited [13]. This highlights a gap in the literature and reinforces the relevance of applying FAIR principles to systems such as I2FVG, which operate in a complex, multi-source, and policy-oriented data environment.

This also suggests that FAIR compliance is not only a technical issue, but also an infrastructural and domain-dependent challenge.

4.2 FAIR in the Context of I2FVG

The challenges and limitations described in the previous chapter highlight several weaknesses and rigidities in the data management of I2FVG. In this context, the FAIR principles provide a suitable framework to address these limitations, particularly those related to data fragmentation, lack of documentation, and limited interoperability, while also supporting improvements in scalability [6, 14].

From the perspective of Findability, assigning persistent identifiers (e.g., URIs) and structured metadata (e.g., JSON files) to each data resource, including provenance information, would allow users to trace each dataset back to its origin. This is particularly important in a complex environment where data undergo multiple transformation steps, requiring clear tracking of both data and processes. Moreover, this principle also applies to data processing scripts, which should be versioned and documented to ensure that each transformation step is traceable.

Accessibility is, to some extent, already addressed, as data access is regulated through clearly defined licenses and permissions. However, improving the documentation of access protocols would further enhance transparency and usability.

Interoperability is a key aspect for the evolution of I2FVG. While some level of internal consistency is already present, the current structure lacks alignment with external standards, vocabularies, and ontologies. Adopting shared and standardized definitions would enable better integration with external data sources and allow data to effectively "speak" a common language [6].

Finally, Reusability directly addresses the lack of clear documentation of system design and data processing workflows. Providing detailed descriptions of data sources, transformation steps, and provenance, together with well-documented and standardized scripts, would significantly improve transparency and reproducibility. Structured FAIRification workflows can support this transition by introducing consistent and reusable processes for data transformation and documentation [7].

By applying these principles, I2FVG can evolve towards a more coherent, maintainable, and scalable data infrastructure.

4.3 Benefits of FAIR Data

The adoption of FAIR principles brings several benefits to complex data environments such as I2FVG.

First, it improves the overall consistency of the system. For example, standardizing data processing workflows (e.g., using a single programming language such as Python) and documenting both data and scripts at each step, together with the use of standardized formats and structured meta-data, reduces ambiguities and facilitates both data integration and script maintenance. In the case of I2FVG, this would also reduce the need for manual intervention in data processing and improve the maintainability of the platform.

Second, FAIR data enhances reproducibility and transparency, allowing both developers and users to better understand how data are generated, processed, and used [1].

Third, interoperability enables integration with external datasets and tools, expanding analytical potential beyond the original domain and supporting more advanced use cases.

Moreover, improved accessibility and documentation reduce dependency on specific developers, facilitating knowledge transfer, long-term maintenance, and scalability.

Finally, FAIR principles increase the overall value of the system, making it more usable, scalable, and aligned with modern data governance practices, particularly in the context of public sector data management [5].

Chapter 5

Methodology and Implementation

5.1 FAIRification Approach

The FAIRification of I2FVG in this work consists of analyzing the components of the system that do not meet FAIR principles and addressing these limitations through a structured and incremental approach.

Since FAIRification is an intensive and resource-demanding process, this work focuses on a limited portion of the overall I2FVG system. The objective is to develop a FAIR-oriented PoC that demonstrates the feasibility of a broader FAIRification process. The PoC includes a demonstrative web-based mock-up interface developed with Django and is intended to validate the technical and methodological feasibility of applying FAIR principles to an existing heterogeneous data infrastructure.

Rather than replacing the original production system, the proposed framework demonstrates how FAIR-oriented workflows, metadata generation, semantic enrichment, and reproducible infrastructures can be progressively integrated into the I2FVG environment.

It is important to note that the proposed approach does not consist of a complete FAIRification of the original I2FVG infrastructure. Rather than fully redesigning the existing database schema, the implementation adopts a FAIR-by-design strategy for newly developed data pipelines while preserving compatibility with the current production environment.

For this reason, the original column names and data structures were intentionally maintained in order to ensure consistency with the existing database and ingestion logic. Within this context, the FAIRification effort mainly concerns metadata generation, provenance tracking, semantic enrichment, and ontology-based crosswalk mechanisms, which enable heterogeneous datasets to be progressively aligned with external semantic standards without requir-

ing disruptive modifications to the underlying infrastructure.

Starting from the available data sources, the analysis concentrates on a selected set of datasets that are representative of the overall infrastructure. In particular, the datasets considered include company registry data, financial information, and EU-funded projects.

Based on these datasets, a dedicated data pipeline has been developed to process, transform, and semantically enrich data using DCAT-compliant metadata structures¹ and ontology-based semantic crosswalks.

Furthermore, a web application has been developed using the Django framework, supported by a SQLite database. The entire system has been containerized using Docker, ensuring portability and reproducibility.

The FAIR-oriented mockup developed in this work is publicly available through a GitHub repository². In addition, an archived release of the project has been published on Zenodo with a persistent DOI³, improving reproducibility and long-term accessibility of the implementation.

The overall architecture of the FAIR-oriented mockup is illustrated in Figure 5.1.

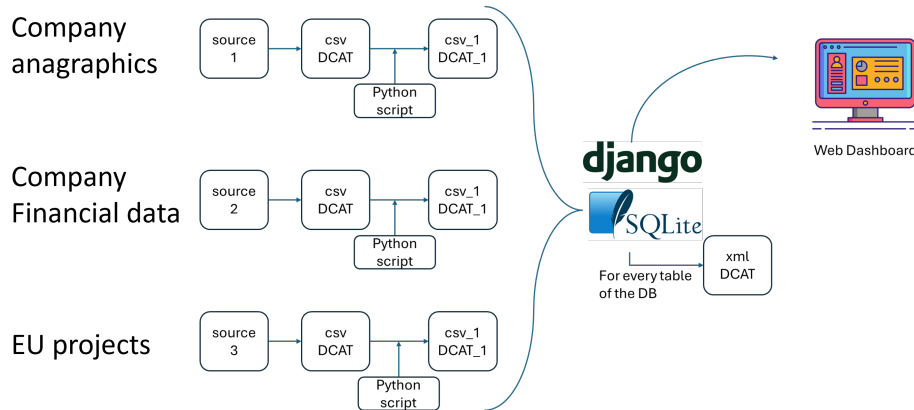


Figure 5.1: General architecture of the FAIR-oriented mockup for I2FVG. Different data domains are processed through dedicated Python-based pipelines that generate structured datasets, DCAT-compliant metadata, and semantic crosswalk mappings before integration into a centralized system based on Django and SQLite.

The methodology follows a modular and domain-oriented FAIRification strategy, in which each component of the system is independently analyzed

¹<https://www.w3.org/TR/vocab-dcat-3/>

²https://github.com/Enrico-Longato/i2fvg_fair_mockup

³<https://zenodo.org/records/20177633>

and progressively redesigned according to FAIR principles.

Each pipeline follows a similar sequence of steps, including data extraction, cleaning, transformation, and metadata generation. In particular, the processing phase includes the creation of DCAT-compliant metadata and semantic enrichment through ontology-based crosswalk mechanisms.

The outputs of the different pipelines are then integrated into a centralized system, implemented through a Django web application supported by a SQLite database. This architecture enables unified access to heterogeneous datasets and facilitates their exploration through a web interface.

The implementation mainly focuses on improving traceability of data transformations, standardizing data processing workflows, introducing structured machine-readable metadata, and partially aligning datasets through ontology-based semantic crosswalks.

Rather than transforming I2FVG into a fully FAIR-compliant infrastructure, the objective is to provide a FAIR-oriented mockup that demonstrates how FAIR principles can be applied to an existing, non-FAIR system. The approach focuses on restructuring key components—data processing, metadata, and system organization—while preserving the original logic of the platform.

This approach allows the system to evolve towards FAIR compliance while acknowledging the infrastructural constraints that prevent full implementation, such as the absence of persistent identifiers and standardized access endpoints.

5.2 Data Processing and Cleaning

The data processing and cleaning phase represents a central step in the FAIR-oriented mockup, as the quality, consistency, and structure of the data directly affect their reusability and interoperability. Since the selected data sources differ in format, origin, and level of standardization, a dedicated preprocessing workflow was designed for each dataset, while maintaining a shared methodological structure across the entire pipeline.

The process begins with data extraction from the original sources. In this work, the selected sources include company registry data, financial data, and information on EU-funded projects. These datasets are characterized by heterogeneous schemas and varying levels of completeness, requiring an initial harmonization effort before they can be integrated into a common system.

After extraction, each dataset undergoes a cleaning phase aimed at reducing inconsistencies and improving overall data quality. This phase includes

the removal of duplicated records, the correction of malformed values, the normalization of textual fields, and the handling of missing or incomplete information. Particular attention is given to identifiers, organization names, dates, and location-related attributes, as these elements are essential for linking records across datasets and supporting subsequent metadata generation.

A transformation step follows the cleaning process. Here, source-specific fields are converted into a more consistent internal representation, enabling datasets with different structures to be processed through a unified logic. This includes schema alignment, attribute renaming, data type normalization, and the conversion of values into machine-readable formats. The objective is not to remove the heterogeneity of the original data sources, but to make it manageable and compatible with the FAIR-oriented architecture of the mockup.

To improve reproducibility, the processing workflow has been implemented through explicit and modular scripts, so that each transformation step can be traced and repeated. This design enhances transparency in data handling and makes the pipeline easier to inspect, maintain, and extend. In this sense, the cleaning process is not only a technical necessity, but also a methodological choice aligned with FAIR principles, particularly with respect to reusability and provenance.

The detailed workflow of the FAIR-oriented data pipeline is illustrated in Figure 5.2.

The outputs of this phase are structured datasets that are sufficiently standardized to support subsequent stages, including metadata generation, semantic mapping, and integration into the centralized application. Although the cleaning procedures do not guarantee full semantic consistency across all sources, they establish a necessary baseline for creating a more coherent and interoperable data environment.

The data processing and cleaning phase demonstrates how even a partial FAIRification effort requires careful intervention at the level of raw data preparation. By introducing systematic cleaning, normalization, and transformation procedures, the mockup reduces fragmentation in the original workflows and creates the conditions for more reliable metadata production and dataset integration.

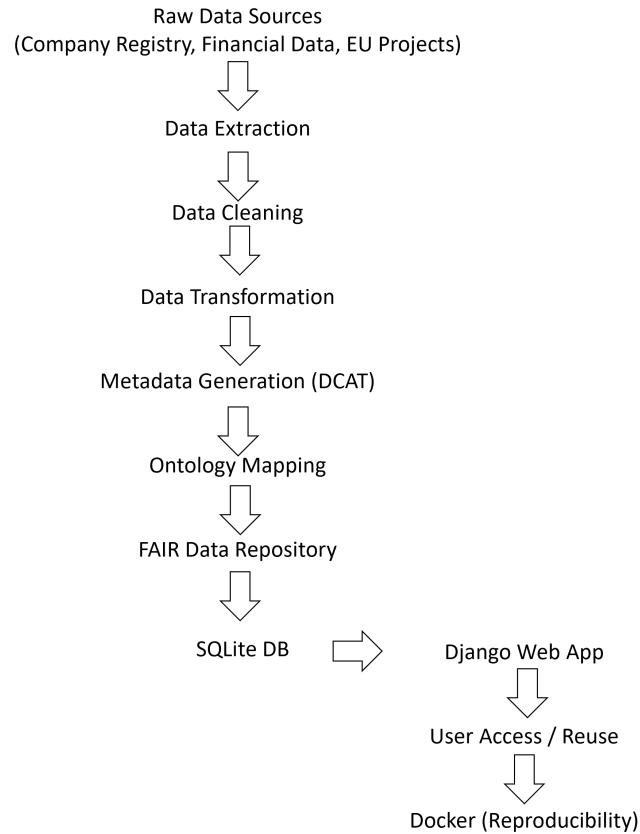


Figure 5.2: Workflow of the FAIR-oriented mockup for I2FVG. The process starts from heterogeneous data sources and proceeds through data extraction, cleaning, transformation, metadata generation (DCAT), and semantic mapping. The resulting datasets are integrated into a FAIR-oriented repository and made accessible through a Django web application supported by a SQLite database. The system is containerized using Docker to ensure reproducibility.

5.3 Documentation and Metadata

A central element of the FAIRification process is the introduction of structured documentation and metadata. In the original I2FVG system, documentation was limited and often implicit, making it difficult to understand data sources, transformations, and dependencies. To overcome this limitation, a metadata-driven approach based on DCAT-compliant JSON structures was adopted.

Each dataset generated within the mockup is associated with a metadata file describing:

- dataset identification information (e.g., title, description, and keywords);
- temporal metadata (creation and modification dates);
- distribution details (format, size, and access path);
- structural information (columns, data types, row and column counts);
- semantic annotations linking attributes to external vocabularies and ontologies;
- provenance information describing the origin of the data and the processing scripts involved in their generation.

Metadata are generated automatically as part of the data processing pipeline and stored in machine-readable JSON files. This ensures consistency between datasets and metadata while reducing the need for manual documentation.

The adoption of the DCAT standard⁴ enables the datasets to be described using a structure aligned with widely adopted metadata practices in data catalogues and public sector environments. In addition, provenance-related information is partially modeled using concepts inspired by the PROV ontology, improving traceability of transformations and processing activities.

A semantic crosswalk mechanism has also been introduced to connect selected attributes to external ontologies and controlled vocabularies. This component represents one of the central contributions of the FAIR-oriented mockup, since it enables heterogeneous datasets to be partially aligned with shared semantic standards while preserving domain-specific characteristics.

⁴<https://www.w3.org/TR/vocab-dcat-3/>

The semantic mappings improve interoperability by associating local attributes with widely adopted vocabularies and ontological concepts. In particular, organizational information, geographical entities, financial indicators, and project-related data were semantically linked to external standards such as Schema.org, SKOS, Registered Organization Vocabulary (ROV), and IFRS Taxonomy.

Although the mapping process remains partial and experimental, it demonstrates how semantic enrichment strategies can support the integration of heterogeneous territorial and innovation-related datasets within FAIR-oriented infrastructures.

The semantic crosswalks were developed independently for different data domains, including company registry information, EU-funded projects, and financial datasets. This approach enabled heterogeneous data sources to be partially aligned with external semantic standards while preserving domain-specific characteristics and supporting interoperability across different domains.

Table 5.1 shows representative examples of the semantic mappings implemented within the FAIR-oriented mockup.

An example of DCAT-based metadata generated within the mockup is shown in Figure 5.3.

```
{
  "@type": "dcat:Dataset",
  "dcterms:title": "i2fvg_company_registry",
  "dcterms:description": "Company registry dataset",
  "dcat:keyword": ["innovation-intelligence"],
  "dcat:distribution": {
    "dcat:mediaType": "text/csv"
  },
  "provenance": {
    "generated_by": "01_02_company_registry.py"
  }
}
```

Figure 5.3: Simplified example of DCAT-based metadata generated for a dataset within the FAIR-oriented mockup.

The ontologies adopted in the mockup are summarized in Table 5.2.

The metadata model adopted in this work does not fully satisfy all FAIR requirements, particularly with respect to persistent identifiers and publicly resolvable access endpoints. Nevertheless, it provides a structured foundation

Domain	Dataset	tribute	At-	Semantic Mapping	Vocabulary	/
Company Reg- istry	piva			schema:vatID	schema.org	
Company Reg- istry	denominazione			rov:legalName	Registered Organi- zation Vocabulary (ROV)	
Company Reg- istry	prov			clv:hasProvince	Core Location Vocabulary (CLV Italy)	
EU Projects	projectID			schema:identifier	schema.org	
EU Projects	totalCost			schema:amount	schema.org	
EU Projects	topics			schema:about	schema.org	+
EU Projects	role			org:role	W3C Organization Ontology	
Financial Data	Totale attivo			ifrs:Assets	IFRS Taxonomy	
Financial Data	Totale patrimonio netto			ifrs:Equity	IFRS Taxonomy	
Financial Data	Utile/perdita eser- cizio ultimi			ifrs:ProfitLoss	IFRS Taxonomy	

Table 5.1: Examples of semantic crosswalks implemented within the FAIR-oriented mockup to align heterogeneous datasets with external vocabularies and ontologies.

Prefix	Ontology / Vocabulary	Purpose in the Mockup
dcat	DCAT	Dataset description, distributions, and catalog metadata generation.
dcterms	Dublin Core Terms	General metadata fields such as title, description, creation date, and modification date.
prov	PROV Ontology	Provenance tracking and description of data generation workflows.
schema	Schema.org	Representation of organizations, addresses, identifiers, dates, and contact information.
skos	SKOS	Representation of controlled vocabularies, classifications, and semantic concepts.
org	Organization Ontology	Description of organizations, organizational sites, and related structures.
rov	Registered Organization Vocabulary	Representation of legal organizations and business-related attributes.
adms	ADMS	Administrative identifiers and metadata associated with public-sector resources.
clv	Core Location Vocabulary	Standardized representation of geographical and administrative locations.
cov	Core Public Organization Vocabulary	Representation of administrative and registry-related identifiers.
ifrs	IFRS Taxonomy	Semantic representation of financial statements and accounting-related concepts.

Table 5.2: Main ontologies and vocabularies used within the FAIR-oriented mockup for semantic annotation and metadata generation.

for improving findability, interoperability, and reusability within the context of the mockup.

Generally speaking, the introduction of structured metadata and semantic annotations significantly improves transparency and traceability across the system. By documenting datasets, transformations, and processing logic in a machine-readable form, the FAIR-oriented mockup reduces ambiguity and facilitates future extensions, maintenance activities, and potential integration with external systems.

5.4 System Simplification and Design

The FAIRification process also involved a simplification and partial redesign of the system architecture.

The original system was characterized by a high level of complexity, including multiple dependencies, rigid components, and strong coupling with a specific technical environment. This limited flexibility, maintainability, and portability.

To address these issues, a simplified architecture has been developed as part of the FAIR-oriented mockup. The redesigned system is based on:

- a lightweight relational database (SQLite) for simplified data storage;
- a web application developed with the Django framework;
- a modular directory structure organized by data domain;
- a containerized environment to ensure portability and reproducibility.

This architecture clearly separates the different layers of the system:

- raw data sources;
- processing pipelines;
- metadata layer;
- application layer.

Such separation improves system clarity, reduces dependencies, and facilitates future extensions.

The adoption of containerization further enhances reproducibility by enabling the system to be deployed consistently across different environments, reducing reliance on specific operating systems.

To facilitate dataset exploration and provide unified access to heterogeneous information sources, a web-based dashboard was developed using the Django framework. The interface integrates company registry data, financial indicators, and EU-funded project information within a centralized environment, demonstrating the practical integration of the FAIR-oriented pipelines, metadata structures, and semantic enrichment mechanisms implemented in the mockup.

The dashboard also demonstrates how semantically enriched heterogeneous datasets can be integrated into a unified exploration environment supporting territorial and innovation-related analysis.

An overview of the developed dashboard is shown in Figure 5.4.

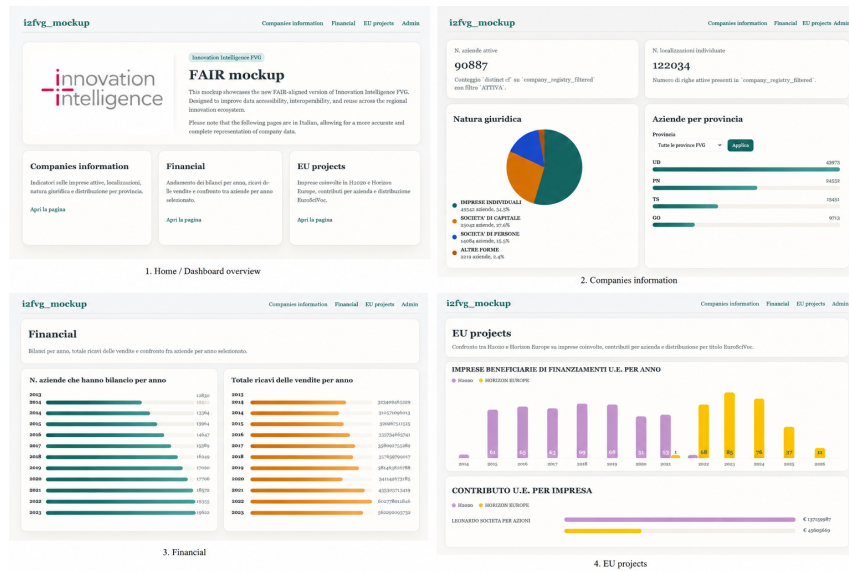


Figure 5.4: Overview of the FAIR-oriented dashboard developed for the I2FVG mockup. The web application integrates multiple data domains, including company registry information, financial datasets, and EU-funded projects, providing unified access through interactive visualizations and aggregated indicators.

The dashboard represents a simplified demonstration layer of the FAIR-oriented mockup, showing how heterogeneous datasets can be explored through a unified and more accessible interface.

The development and testing activities were conducted within a virtual machine hosted on the ORFEO high-performance computing (HPC) infrastructure provided by Area Science Park⁵. This environment enabled the exe-

⁵<https://orfeo-doc.areasciencepark.it/>

cution of data processing workflows without overloading local computational resources and provided a controlled environment for dependency management.

A dedicated Python virtual environment was configured within the OR-FEO virtual machine and later encapsulated through Docker containerization, improving portability and reproducibility of the FAIR-oriented mockup across different systems.

It is important to note that this redesigned architecture does not aim to replace the original I2FVG infrastructure, but rather to provide a simplified and FAIR-oriented reference model that demonstrates how the system could evolve towards more modern and interoperable data management practices.

Chapter 6

Results and Discussion

6.1 Improvements Achieved

The project demonstrates the feasibility of applying FAIR principles to I2FVG, representing a first step towards the development of a more FAIR-oriented data infrastructure.

Although the proposed implementation does not aim to achieve full FAIR compliance, the FAIR-oriented mockup introduces several improvements compared to the original system architecture and data management approach.

First, the project improves the overall organization of the system by introducing a more modular architecture. Data processing workflows are organized into independent pipelines structured according to their respective domains and coordinated through a central controller. This approach simplifies the management of data preparation steps, improves maintainability, and facilitates the future integration of additional data sources.

Second, the introduction of structured metadata generation significantly improves documentation and traceability. Through the adoption of DCAT-compliant JSON metadata, datasets become more understandable, machine-readable, and easier to manage. In addition, metadata generation is integrated directly into the processing pipeline, enabling transformations and provenance-related information to be automatically documented.

Another important improvement concerns interoperability, which represents a fundamental step towards openness and integration with external systems. The implementation of semantic mappings and ontology crosswalks partially aligns the datasets with external vocabularies and standardized concepts, facilitating integration with external infrastructures and improving semantic consistency.

The FAIR-oriented mockup also improves reproducibility and portability.

The use of modular scripts, Docker containerization, and a public GitHub repository supports transparency and enables the system to be more easily replicated, transferred, and extended across different computational environments.

Furthermore, the redesigned architecture simplifies several aspects of the original infrastructure. The use of a lightweight SQLite database and a centralized Django web application reduces system complexity while still supporting unified access to heterogeneous datasets.

Table 6.1 summarizes the main differences between the original I2FVG infrastructure and the FAIR-oriented mockup developed in this work.

The evaluation highlights how interoperability and metadata traceability represent the most mature aspects of the proposed FAIRification strategy, while accessibility and findability still require additional infrastructural developments.

Aspect	Original System	FAIR-oriented Mockup
Metadata management	Limited and partially implicit	Automated DCAT-compliant metadata generation
Data processing workflows	Fragmented and heterogeneous	Modular domain-oriented pipelines
Semantic interoperability	Absent or limited	Ontology-based semantic crosswalks
Reproducibility	Environment-dependent	Docker-based reproducible infrastructure
Documentation	Partial and manual	Structured and machine-readable
System architecture	Strongly coupled components	Simplified layered architecture
Dataset integration	Difficult to extend	Easier integration of additional data sources

Table 6.1: Comparison between the original I2FVG infrastructure and the FAIR-oriented mockup.

In general these improvements demonstrate how FAIR principles can support the evolution of complex and heterogeneous data systems towards more transparent, interoperable, maintainable, and reusable infrastructures.

6.2 Critical Evaluation of the FAIR-oriented Approach

Although the FAIR-oriented mockup introduces several improvements in terms of metadata management, interoperability, and reproducibility, the implementation only partially satisfies the FAIR principles. This is mainly due to the demonstrative nature of the mockup itself, which was designed as a PoC rather than as a fully operational FAIR-compliant infrastructure.

As a consequence, some FAIR-related components were implemented only partially and at a limited scale. In addition, the original I2FVG system is characterized by a complex and heterogeneous architecture that would require a broader infrastructural redesign and a significantly larger implementation effort to fully satisfy all FAIR requirements.

Therefore, the level of compliance varies across the four FAIR dimensions, depending both on the maturity of the implemented components and on the infrastructural constraints inherited from the original system.

In addition, aspects related to authentication, authorization, and access management were not implemented in the mockup, since the project was mainly designed as a PoC focused on metadata management, interoperability, and reproducibility.

Table 6.2 summarizes the extent to which the proposed mockup addresses the FAIR principles.

The evaluation highlights how the strongest improvements concern interoperability and metadata standardization, mainly due to the adoption of semantic vocabularies and structured metadata models. At the same time, several limitations remain unresolved, particularly regarding persistent identifiers, machine-actionable access mechanisms, and fully standardized licensing information.

These limitations are largely connected to infrastructural constraints and to the demonstrative nature of the mockup itself, which was designed as a FAIR-oriented PoC rather than a production-ready FAIR data infrastructure.

Principle	Guideline	Application to I2FVG
Findable	Use of persistent identifiers and rich metadata to enable data discovery	Partial implementation. Datasets are described through structured DCAT-based metadata, including titles, descriptions, and keywords. However, persistent and globally resolvable identifiers are not implemented, limiting full compliance.
Accessible	Data retrievable via standardized communication protocols with defined access conditions	Partially implemented. The mockup provides access to datasets through a web-based interface and machine-readable metadata structures. However, persistent identifiers, APIs, and standardized machine-accessible endpoints are not currently implemented.
Interoperable	Use of shared standards, formats, and vocabularies to enable data integration	Good implementation. Data are managed through structured formats such as CSV files and a relational SQLite database, while metadata are generated in machine-readable JSON/DCAT structures and enriched with semantic annotations using standard vocabularies (e.g., DCAT, schema.org, SKOS). Some inconsistencies remain in vocabulary alignment and metadata completeness.
Reusable	Rich metadata, clear provenance, and usage conditions to support data reuse	Partially implemented. Detailed metadata and provenance information are provided, including data sources and processing steps. However, licensing is not consistently expressed in machine-readable form and documentation can be further improved.

Table 6.2: Application of FAIR principles to the I2FVG system

Chapter 7

Conclusions and Future Developments

7.1 Conclusions

This work presented a FAIR-oriented redesign of the Innovation Intelligence FVG (I2FVG) system through the development of a demonstrative mockup aimed at improving data organization, interoperability, reproducibility, and metadata management.

Although I2FVG already represents a valuable example of data integration for territorial analysis, particularly in the context of entrepreneurship and innovation capacity assessment, the original infrastructure presents several limitations related to data processing workflows, metadata management, interoperability, and system maintainability. These limitations motivated the development of a more structured, scalable, and FAIR-oriented approach.

Starting from the analysis of the original infrastructure and its criticalities, the thesis explored how FAIR principles can be incrementally applied to an existing heterogeneous system that was not originally designed according to FAIR methodologies. In particular, the work focused on restructuring data processing workflows, introducing automated metadata generation mechanisms, implementing semantic crosswalks based on external ontologies, and simplifying the overall system architecture.

The proposed implementation combines modular Python-based data pipelines, DCAT-compliant metadata generation, semantic enrichment through ontology mappings, a Django-based web application, and Docker containerization. The resulting mockup highlights the potential of FAIR-oriented approaches for improving transparency, maintainability, and interoperability in complex data infrastructures

Although the implementation does not achieve full FAIR compliance, the work highlights how even a partial FAIRification effort can produce significant improvements in terms of metadata quality, traceability, semantic consistency, and reproducibility. At the same time, the thesis emphasizes the practical challenges associated with FAIRification processes, particularly when dealing with complex legacy systems characterized by heterogeneous data sources and infrastructural constraints.

The developed mockup should therefore be considered both as a proof-of-concept (PoC) and as a possible reference for future FAIR-oriented evolutions of I2FVG and similar data-intensive systems.

7.2 Future Developments

The FAIR-oriented mockup developed in this work represents a first important step towards a broader and more structured evolution of the I2FVG infrastructure. Although the current implementation mainly serves as a demonstrative PoC, the results highlight how existing data-intensive systems can be progressively re-engineered according to FAIR principles in order to improve transparency, usability, interoperability, data sovereignty, and long-term sustainability.

The proposed approach is particularly relevant for research and public-sector environments characterized by heterogeneous, evolving, and partially unstructured datasets, where interoperability and reproducibility are becoming increasingly important requirements for both analytical and policy-making activities.

Several future developments may extend and strengthen the FAIR-oriented approach presented in this thesis.

A first important direction concerns the integration of Large Language Models (LLMs) to support the standardization and semantic interpretation of heterogeneous and non-tabular data sources. In particular, LLM-based approaches could facilitate metadata extraction, semantic normalization, and ontology alignment processes, especially when dealing with partially unstructured textual information. However, such approaches should be integrated within human-in-the-loop validation workflows in order to preserve data quality, transparency, and semantic consistency.

Another relevant development involves the implementation of machine-accessible APIs to enable programmatic access to datasets, metadata, and analytical services. This would significantly improve interoperability with external systems and support integration with other FAIR-oriented infrastructures.

Future work may also extend the metadata layer by adopting additional semantic standards and provenance models, further improving traceability, semantic interoperability, and metadata richness.

From an infrastructural perspective, the current SQLite-based implementation could evolve towards more scalable database solutions, such as PostgreSQL, in order to better support larger datasets, concurrent access, and more advanced analytical functionalities.

Overall, the developed system should be considered a FAIR-oriented PoC supported by a demonstrative mock-up interface and a modular FAIRification methodology.

More broadly, the work demonstrates how incremental FAIRification strategies can provide a realistic pathway for evolving complex legacy infrastructures toward more interoperable, transparent, and reusable data ecosystems without requiring complete infrastructural replacement.

Bibliography

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, “The fair guiding principles for scientific data management and stewardship,” *Scientific Data* **3** no. 1, (March, 2016) 160018.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC4792175/>.
- [2] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, “Benefits, adoption barriers and myths of open data and open government,” *Information Systems Management* **29** no. 4, (October, 2012) 258–268.
<https://doi.org/10.1080/10580530.2012.716740>.
- [3] S. Badiee, J. Crowell, L. Noe, A. Pittman, C. Rudow, and E. Swanson, “Open data for official statistics: History, principles, and implementation,” *Statistical Journal of the IAOS* **37** no. 1, (January, 2021) 139–159. <https://opendatawatch.com/wp-content/uploads/2021/Publications/Open-Data-for-Official-Statistics-History-Principles-Implementation.pdf>.
- [4] B. Ubaldi, “Open government data: Towards empirical analysis of open government data initiatives,” *OECD Working Papers on Public Governance* no. 22, (May, 2013) 1–60.
https://interoperable-europe.ec.europa.eu/sites/default/files/document/2015-08/oecd_barbara_ubaldi.pdf.

- [5] European Commission Expert Group on FAIR Data, “Turning fair into reality: Final report and action plan from the european commission expert group on fair data,” November, 2018.
<https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1>.
- [6] D. Welter, N. Juty, P. Rocca-Serra, F. Xu, D. Henderson, W. Gu, J. Strubel, R. T. Giessmann, I. Emam, Y. Gadiya, *et al.*, “Fair in action: a flexible framework to guide fairification,” *Scientific Data* **10** (May, 2023) 291. <https://doi.org/10.1038/s41597-023-02167-2>.
- [7] A. Jacobsen, R. Kaliyaperumal, L. B. da Silva Santos, B. Mons, E. Schultes, M. Roos, and M. Thompson, “A generic workflow for the data fairification process,” *Data Intelligence* **2** no. 1-2, (January, 2020) 56–65. https://doi.org/10.1162/dint_a_00028.
- [8] L. Vogt, “The clear principle: organizing data and metadata into semantically meaningful types of fair digital objects to increase their human explorability and cognitive interoperability,” *Journal of Biomedical Semantics* **16** no. 1, (2025) 18. <https://link.springer.com/content/pdf/10.1186/s13326-025-00340-7.pdf>.
- [9] A. Longo, M. Zappatore, D. Colella, F. Zampino, and A. Calò, “xfair: A multi-layer approach to data fairness assessment and data fairification,” *ACM Journal of Data and Information Quality* **17** no. 4, (December, 2025) Article 25.
- [10] B. dos Santos Vieira, C. H. Bernabé, S. Zhang, H. Abaza, N. Benis, A. Cámara, R. Cornet, C. M. A. Le Cornec, P. A. C. 't Hoen, F. Schaefer, K. J. van der Velde, M. A. Swertz, M. D. Wilkinson, A. Jacobsen, and M. Roos, “Towards fairification of sensitive and fragmented rare disease patient data: challenges and solutions in european reference network registries,” *Orphanet Journal of Rare Diseases* **17** no. 1, (2022) 448.
- [11] A. A. Sinaci, M. Gencturk, H. Teoman, G. Laleci Erturkmen, C. Alvarez-Romero, A. Martinez-Garcia, B. Poblador-Plou, J. Carmona-Pérez, M. Löbe, and C. L. Parra-Calderon, “A data transformation methodology to create findable, accessible, interoperable, and reusable health data: Software design, development, and evaluation study,” *Journal of Medical Internet Research* **25** (2023) e42822. <https://www.jmir.org/2023/1/e42822>.

- [12] J. Schaaf, D. Kadioglu, J. Goebel, C.-A. Behrendt, M. Roos, D. van Enckevort, F. Ückert, F. Sadiku, T. O. F. Wagner, and H. Storf, “Osse goes fair – implementation of the fair data principles for an open-source registry for rare diseases,” in *German Medical Data Sciences: A Learning Healthcare System*, U. Hübner and et al., eds., pp. 209–214. IOS Press, 2018.
- [13] P. Potturi, “Demystifying semantic layers in business intelligence platforms,” *Sarcouncil Journal of Engineering and Computer Sciences* 4 no. 8, (August, 2025) .
<https://doi.org/10.5281/zenodo.16886966>. Licensed under CC BY-NC-ND 4.0.
- [14] Y. Gadiya *et al.*, “Fair data journey: from data integration to data reuse,” *Scientific Data* (2023) .