RESILIAGE
Revolutionising community resilience

D1.3
Data Lake model

Project Title	RESILIAGE – Advancing holistic understanding of community RESILIence and cultural natural heritAGE drivers through community-based methodologies
Programme	Horizon Europe
Grant Agreement	101121231
Start of the Project	September 1, 2023
Duration	36 months

Funded by the European Union

Deliverable number	D1.3
Version	1.0
Deliverable type	Document, Report
Actual date of delivery	30/06/2024
Dissemination level	Confidential
Work Package	WP1
Lead beneficiary	POLITO
Main Authors	Rosa Tamborrino, Enrico Macii (POLITO)
Contributors	
Peer Reviews	Luís Manteigas da Cunha (Almende) Fernando Núnez (Vexiza)

VERSION	DATE	PARTNER	MODIFICATIONS
0.1	30/06/2024	POLITO	

Index of Content

Index of tables

Index of figures

Table of Abbreviations and Acronyms

Abbreviation	Meaning
3Vs	Volume, Velocity, and Variety
API	Application Programming Interface
AWS	Amazon Web Services
AWS S3 or S3	Amazon Simple Storage Service
CCA	Climate Change Adaptation
CNH	Cultural Natural Heritage
CIAM	Customer Identity and Access Management
CORE labs	Community Resilience laboratories
CRS	Coordinate Reference System
CSV	Comma-separated Values
D	Deliverable
DDP	Data-Driven Platform
DB	Database
DFS	Distributed File Systems
DMS	Data Management System
DRM	Disaster Risk Management
DRR	Disaster Risk Reduction
DSS	Decision Support System
ETL	Extract/Transform/Load
FAIR	Findable, Accessible, Interoperable and Re-usable
FG	Focus Group

GeoTIFF	Geospatial Tagged Image File Format
GIS	Geographic Information System
HDFS	Hadoop Distributed File System
IAM	Identity and Access Management
IoT	Internet of Things
JSON	JavaScript Object Notation
KLM	Keyhole Markup Language
LL	Lesson learned
OGC	Open Geospatial Consortium
POI	Point Of Interest
RAISE	Resilience Assessment Interactive Self-Enabler tool
RBAC	Role-Based Access Control
SD	Sustainable Development
SDK	Sofware Development Kit
SQL	Structured Query Language
T	Task
TIFF	Tagged Image File Format
VPC	Virtual Private Cloud
XML	eXtensible Markup Language

1. Executive Summary

The Data Lake model is a document of the RESILIAGE project, delivered in the context of WP1, Task 1.3: Data Lake model. The objective of WP1 is the analysis of existing knowledge through literature and policies scoping review for focusing on human behaviour as triggering or cascading factors of disasters or crisis situations and on heritage drivers of community resilience, to improve analyses and transform qualitative data into quantitative information (WP2). In detail, the WP1 aims i) to identify international standards for crisis preparedness planning (to implement the SENDAI Framework) and Climate Change Adaptation (CCA) policies and strategies; ii) determine existing best practices and lessons learned from past experiences to enhance citizens' preparedness and societal resilience in a multi-hazard scenarios in the 5 COmmunity REsilience laboratories (CORE labs) involved, and beyond; iii) analyse the cascading effects of human factors and behavioral aspects in crises and disaster by considering a systemic approach to the Disaster Risk Management (DRM) ; and design a multiscale, multisource data model for the RESILIAGE research project. Within it, Task 1.3 aims to design and implement a model for a multiscale, multisource data lake to manage the extensive amount of data and information that will be extracted, addressed, co-created, disaggregated, structured, shared, operationalized, researched, and exploited in the project. The RESILIAGE Data Lake wants to be a centralized repository to store, use, and reuse both structured data (including external data such as Copernicus data, Galileo data, national and local datasets, new datasets, etc.) and unstructured data such as information collected and created by the knowledge building of the project (including text, images, videos, audio, 3D models, maps, documents, conceptual maps, graphs) at any scale. Moreover, Task 1.3 defines the technical requirements of Findable Accessible Interoperable Reusable (FAIR) data (see also D8.2) and parameters for modelling the identified factors and develops a common glossary of terms. Additionally, it designs the data model to be integrated into the Resources Ecosystem for Community Resilience (WP3). By storing data as-is, without prior structuring, this Task enables various types of analyses (ranging from dashboards and visualisations to big data processing, real-time analytics, and machine learning) to critically guide better decisions. This includes analysing knowledge collected and generated during the project activities, such as information on recent disasters, past disasters, collective memory, local knowledge, socio-demographics, cultural practices, cultural identifications, and socio-cultural-psychological-behavioral factors. The RESILIAGE Data Lake model considers data generated within WP1 and the ongoing activities in other WPs (especially WP2, WP3, WP4, WP5, WP6), and creates the best model to include all data generated by the RESILIAGE project throughout its various WPs to feed WP3 Resource Ecosystem and its integrated digital tools.

The objective of this deliverable is to describe the solutions adopted for Data Lake management in terms of the platform and software used. This report thus contributes to detailing the data management infrastructure of the RESILIAGE project for the scope of RESIALIGE Resource Ecosystem (WP3). It provides the overall approach, methods, and a set of components to ensure high-reliability data storage and efficient information collection. By adopting the most efficient solutions for storing the relevant information collected and produced by RESILIAGE, the report offers technical documentation of the implemented Data Lake and data management system.

The document is organized as follows. Section 2 outlines the RESILIAGE project overview by introducing the rationale for drafting this deliverable. Section 3 provides a comprehensive overview of the enabling technologies and paradigms fundamental to the data lake and related data management. Finally, Section 4 details the implementation solutions for deploying the project data lake. The concluding section of the document encompasses the scope of this deliverable and the bibliography, which lists the references cited throughout the document.

2. Introduction

2.1. Project Overview

Recent crises and disasters have profoundly impacted the lives, livelihoods, and environment of European citizens. RESILIAGE, aligned with the Sendai Framework and the 2030 Agenda, answers to the need to better understanding human behaviours which are influenced by interlinked psychological, social, cultural, historical, economic, and environmental factors. It frames Disaster and Risk Reduction throughout community resilience, and heritage drivers as a resource of local communities.

RESILIAGE aims to advance understanding of community resilience and enhance it, through holistic actionable knowledge generation with local communities in multi-hazard and multidimensional systemic innovation frameworks by exploring heritage as a driver for enhancing community resilience, and contributing to SENDAI framework and sustainable development. It recognizes community practices as "Cultural Natural Heritage" (CNH), accessing it for Disaster Risk Management (DRM) to enhance SENDAI and Climate Change Adaptation (CCA) strategies. RESILIAGE will increase community resilience at local level via a systemic approach, heritage drivers and bottom-up processes to empower local communities.

RESILIAGE addresses this challenge framing community practices and researching, mapping, and exploiting them to advance a holistic understanding of community resilience to support and enhance EU Disaster Risk Reduction policies and action plans, the SENDAI framework, and the EU Climate Adaptation Strategy by: (i) Ensuring active engagement of many citizens and diverse societal groups in the project activities; (ii) Gathering reliable quantitative and qualitative data from diverse local communities in a consistent and reliable manner on how societal factors such as historical, cultural, emotional factors, as well as gender, influence risk perception, communication and resilience (iii) Shaping a collaborative framework and facilitating genuine co-development and co-creation of knowledge with community resilience labs in real scenarios (iv) Demonstrating the benefits and impact of project tools and solutions in 5 CORE Labs across diverse cultural, socio-economic, geographic conditions in a multihazard multiscalar scenario (v) Developing community-centred targeted communication strategies and engaging with relevant stakeholders. In these aims, RESILIAGE explores and experiences digital and non-digital tools, methods, and solutions to co-create better-connected communities with first responders, better support for post-disaster traumas, increased awareness of the physical environment and mindfulness of its heritage drivers.

Due to the information complexity and the diverse data sources, the RESILIAGE framework will be implemented in a multiscale and multisource Data-Driven Platform (DDP), able to provide the necessary information for feeding the RESILIAGE digital platform with integrated digital tools for enhancing community resilience by monitoring and supporting the target groups (namely policy makers, first responders, knowledge organisations, citizen organisations), and replicating the RESILIAGE approach and fundings. All the developments of the project will be validated in 5 CORE labs, representative of main climatic and environmental challenges in cultural regions of Europe and beyond, as detailed in Figure 1.

		CORE labs large-scale scenarios				
		Karşıyaka CORE lab	Trondheim CORE lab	Naturtejo CORE lab	Crete CORE lab	Famenne-Ardenne CORE lab
Affected population		340.000	180.000	86.729	630.000	67.000 ca.
SyRI		Adaptive Governance	Health and Wellbeing	Social interaction and inclusiveness	Active Memory	Socio-economic resilience
Main Hazards	Heatwaves					
	Landslides					
	Earthquakes					
	Wild-fires					
Other Hazards	Floods					
	Rainstorms					
	Urban fires					
Governance scale		City District	Municipality	Municipality network	Regional	Cross-regional
CORE lab network (e.g., citizens associations, first responders, policy makers, vulnerable groups)		✓	✓	✓	✓	✓
Existing tools		IZUM (Izmir Disaster and Transportation Communication Tool)	CIM for alerting volunteers and staff	App for real time information on national wildfires; GIS information App on national Forest wildfires, and fire risk provider	Permanent exhibition hall of NHMC which informs and trains on natural hazards, the Earthquake simulator, the Evande distant learning platform	Geoportal for real data collection for floods hazard, Flood Management Risk plan, Emergency alarm system app, Citizen preparedness website

Figure 1 - Overview of CORE Labs

2.2. Aim of the Report

The objective of this report is to present the implemented solution for effective data management within the context of the RESILIAGE project. At the heart of this endeavour is the detailed explanation and demonstration of the deployed solution, underscored by an extensive survey of the most cutting-edge technologies currently available for data storage and management. This report explores and compares these advanced technologies, shedding light on their capabilities and suitability for the project's requirements. It identifies the data lake model more suitable for the RESILIAGE data, projected tools in the RESILIAGE Resource Ecosystem, and the overall objectives.

To begin with, the report provides an exhaustive introduction to all the essential components of the data management solution. A key focus is placed on describing the concept of a Data Lake, detailing its structure, functionality, and the significant benefits it offers. Emphasis is placed on the advantages of implementing an in-cloud solution compared to traditional on-premises data storage systems. This comparative analysis highlights the superior scalability, flexibility, and cost-effectiveness of cloud-based solutions, which are critical for managing the large and diverse datasets associated with the RESILIAGE project.

Further, the report delves into the concept of the data catalogue, an indispensable tool for efficient data management. It introduces various existing solutions, providing a thorough analysis of their features and functionalities. Moreover, this report includes a discussion on how these solutions facilitate the organisation, discovery, and governance of data. Regardless of whether the RESILIAGE Data Lake stores structured or unstructured data, the report emphasizes the importance of predefined metadata. This metadata is crucial as it encapsulates the essence of the stored information, thereby enabling a streamlined and agile search process. By ensuring that metadata is

meticulously defined and implemented, the report demonstrates how the search and retrieval of data can be significantly optimized.

In summary, this report makes a substantial contribution to the understanding of the data management infrastructure employed in the RESILIAGE project. It documents a suite of components and methodologies designed to guarantee a highly reliable service for data storage and the efficient collection of information. By adopting the most effective solutions for data storage and management, the report provides an in-depth technical documentation of the implemented Data Lake and the associated data management processes. This documentation not only enhances comprehension but also ensures the robustness and reliability of the data management solution, thereby supporting the overarching goals of the RESILIAGE project.

3. Introduction to Data Lake

This section introduces and discusses the various technologies and paradigms that form the backbone of the design and implementation of the RESILIAGE Data Lake. It delves into the details of the advanced technologies that have been carefully selected to ensure the Data Lake meets the stringent requirements of scalability, flexibility, and efficiency.

3.1. Big Data

Nowadays, the term "Big Data" has emerged as a cornerstone concept in the realm of data management and analysis. Big Data refers to the huge volumes of structured and unstructured data generated at high velocity from a multitude of sources, ranging from social media and sensors to transactional systems and more (SAGIROGLU & SINANC, 2013).

Big Data is often described through the lens of the "3Vs" model, which encapsulates its core attributes: Volume, Velocity, and Variety:

- **Volume:** The sheer amount of data being generated is staggering. Gigabytes, terabytes, and even zettabytes (trillions of gigabytes) are not uncommon. This data originates from a multitude of sources, including social media activity, sensor networks, and financial transactions.
- **Variety:** Big data goes beyond the traditional structured data found in Databases (DBs). It encompasses unstructured data like text, images, videos, and social media posts. This variety presents a challenge in terms of analysis and interpretation.
- **Velocity:** The speed at which data is generated and needs to be processed is another key aspect. Real-time data streams from financial markets or social media feeds require near-instantaneous analysis to extract valuable insights.

Beyond the 3Vs, other characteristics like Veracity (the uncertainty of data) and Value (the insights derived from data) are also considered, emphasizing the complexity and potential of Big Data.

How to critically and theoretically analyse Big Data in culture and society is a cogent matter to be considered in relation to other elements such as digital culture and digital humanities methodologies output as well as traditional disciplines outputs. Scholars question the impact of digitised data on knowledge building in relation to complex data that are the most likely data type that human activity tends to produce. (Nugent-Folan, 2020)

The processing of Big Data in culture and research world has also fostered European Commission recommendations for providing critical approach to digital media and information literacy in society to create awareness about of the processing of cultural Big Data. (DEPUTIES, 2017)

The management of big data necessitates specialized tools and techniques (Tsai, 2015). At the heart of managing massive datasets are distributed processing systems. These systems are designed to handle enormous volumes of data by distributing the workload across multiple computers or nodes in a network. This distributed approach not only

enhances the processing power but also ensures fault tolerance and scalability. Unlocking the hidden patterns and insights within Big Data is where advanced analytics techniques come into play. These techniques go beyond traditional data analysis, employing sophisticated algorithms and computational methods to extract meaningful information. Keys among these techniques are: machine learning and artificial intelligence. Before analysis can begin, Big Data must be efficiently ingested and integrated from various sources. This involves capturing data from multiple streams, including transactional DBs, sensor networks, social media platforms, and more. Effective data storage solutions are critical for managing Big Data. Traditional relational DBs often fall short in this regard due to their limitations in handling large volumes and varied data types. Instead, NoSQL DBs, such as MongoDB, Cassandra, and HBase, offer flexible schema designs and horizontal scalability, making them well-suited for Big Data applications. Additionally, cloud storage solutions, such as Amazon S3, Google Cloud Storage, and Azure Blob Storage, provide scalable and cost-effective options for storing vast amounts of data.

The potential benefits of Big Data are vast. Businesses can use it to gain a deeper understanding of their customers, optimize operations, and develop innovative products and services. However, ethical considerations around data privacy and security remain paramount as Big Data continues to evolve.

3.2. Data Catalogue

The RESILIAGE Data Lake takes into account all relevant rpincipia and methods for data cataloguing. Data cataloguing refers to the systematic process of creating an organized inventory of an organization's data assets (Oliveira, 2024). Data cataloguing aims to bring order to the vast and often chaotic world of organisational data. It achieves this by leveraging metadata (described hereafter).

This metadata serves as a detailed description of each data asset, including information like:

- **Data type:** Structured (e.g., tables in databases) or unstructured (e.g., text documents, emails);
- **Content description:** A summary of what the data represents (e.g., customer information, financial records, sensor readings);
- **Location:** Where the data is physically stored within the organization's systems;
- **Lineage:** The origin and transformation history of the data;
- **Access controls:** Who is authorized to access and use the data.

As a well-maintained data catalogue, RESILIAGE provides a central repository for these metadata, allowing users to easily discover, understand, and locate the data they need. This translates into several key benefits:

- **Improved data accessibility:** Users can quickly find relevant data, eliminating wasted time searching through scattered sources;
- **Enhanced data quality:** By documenting data lineage and content, the catalog helps identify and rectify errors or inconsistencies;
- **Better data governance:** Clearer understanding of data ownership and access controls aids in data security and regulatory compliance;

- **Increased data utilisation:** Users are more likely to leverage data they can easily discover and understand, fostering data-driven decision making.

Data cataloguing tools are flexible to range from simple spreadsheets to sophisticated software solutions. The ideal approach depends on the organisation's size, data complexity, and specific needs. Regardless of the chosen method, effective data cataloguing approach is pursued as a crucial step towards maximizing the value of an organisation's ever-growing data assets.

3.3. Metadata

Metadata identifies essentially “data about data”. It provides information about other data, helping to identify, describe, and manage it effectively (Zeng, 2020). Metadata is used to facilitate the understanding, use, and management of data.

In essence, metadata is a structured way of describing characteristics, attributes, and context of an information. This hidden language behind data allows for efficient organisation, discovery, and utilization as depicted in Figure 2.

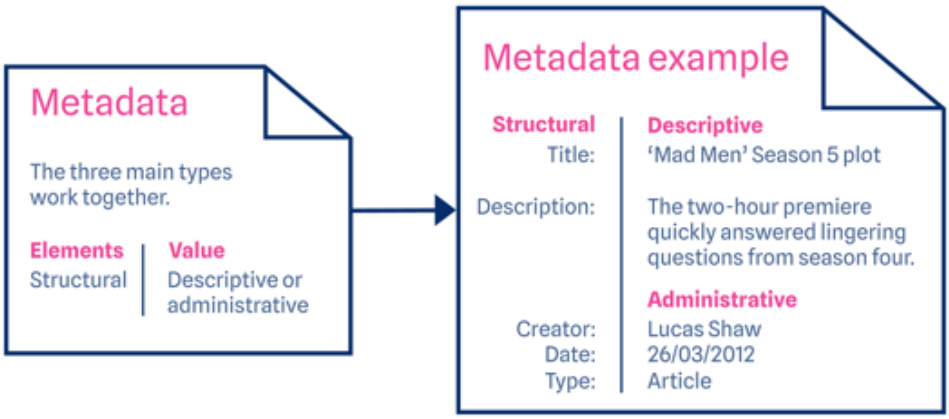

Figure 2 - Metadata Conceptualization

There are various types of metadata, each serving a specific purpose:

- **Descriptive metadata** provides a general summary of the data content, such as title, author, keywords, or creation date;
- **Technical metadata** focuses on the technical details of the data format, file size, or encoding standards;
- **Structural metadata** describes how the data is organized within a file or database, detailing relationships between different elements;
- **Administrative metadata** tracks information related to data ownership, access rights, and usage restrictions.

Metadata plays a crucial role in various aspects of data management. It helps users find relevant data quickly and accurately by providing searchable attributes. It aids in organising and categorising data, making it easier to navigate and understand large datasets. Metadata also helps in combining data from different sources by providing

context and structure. Administrative metadata is essential for data preservation, ensuring that data remains accessible and usable over time. Metadata supports data governance by providing information about data lineage, usage, and compliance with regulations.

The benefits of well-defined metadata are several:

- **Enhanced discoverability:** Users can easily locate relevant data based on its descriptive characteristics;
- **Improved data understanding:** Metadata clarifies the meaning and context of the data, fostering better interpretation;
- **Streamlined data sharing:** Standardised metadata allows for seamless exchange of data between different systems and users;
- **Boosted data quality:** Clear documentation helps identify and rectify errors or inconsistencies within the data;
- **Efficient data governance:** Metadata facilitates data security and regulatory compliance by clarifying ownership and access controls.

Effective metadata creation requires careful consideration of data types, intended uses, and potential users. Standardised metadata schemas, like Dublin Core or ISO standards, can ensure consistency and interoperability across different systems.

In RESILIAGE metadata are not an optional extra. It is critical component of data management, providing essential information that enhances the usability, organisation, and preservation of data. It acts as an enabler for efficient data handling and maximizes the value of the data by making it more accessible and understandable by unlocking the true value of data, transforming it from a disorganised collection into a powerful asset for informed decision-making.

3.4. Geospatial data

RESILIAGE approach and tools also include georeferenced information. The RESILIAGE Multidimensional Atlas for Community Resilience, in particular, will include spatialised information (Task 3.3). Geospatial data includes a wide array of data types, ranging from natural features of CORE labs, such as natural areas and elements in Geoparks to human-made structures including built elements such as buildings and infrastructures as well as intangible cultural natural heritage elements that also can be linked to space.

Geospatial data, also referred to as geographic data or spatial data, encompasses any information that has a direct or indirect reference to a location on Earth's surface. This data is inherently tied to geographic coordinates, typically latitude and longitude, which allow each data point to be precisely located on a map or within a geographic information system (GIS) (Guo, 2020). Key aspects of geospatial data include its spatial component and attributes. The spatial component ensures that each data point has a geographic reference, enabling spatial analysis and visualisation. Attributes describe the characteristics of the features at specific locations.

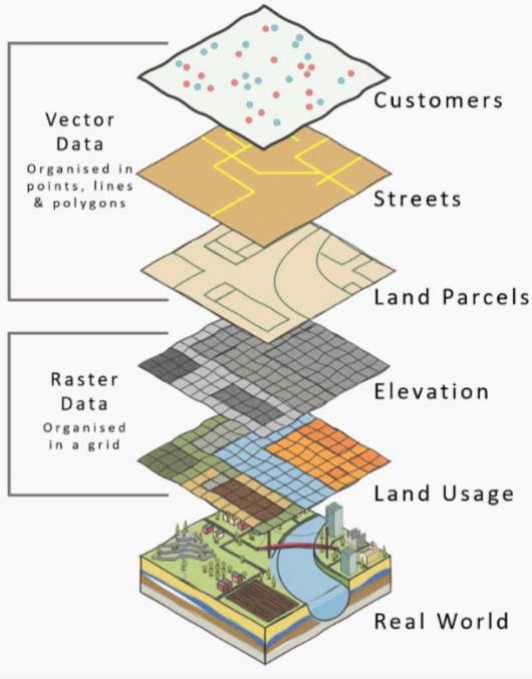

Figure 3 - GIS Conceptualization

For example, attributes include the elevation of a mountain in Crete CORE lab, the type of soil in a region of Naturtejo Geopark in Portugal, or the population density of a municipality such as Karsyaka in Turkey. This data can come in various forms, including:

- **Points of Interest (POI):** Representing specific locations like buildings, parks, weather stations, or historical landmarks. These are typically defined by latitude and longitude coordinates.
- **Lines:** Depicting linear features like roads, rivers, pipelines, or flight paths. Line data is often represented as a sequence of connected points.
- **Polygons:** Defining areas like city boundaries, land parcels, environmental zones, or administrative regions. Polygons are formed by closed loops of coordinates.
- **Satellite Imagery:** Captured from orbiting satellites, providing high-resolution visuals of Earth's surface, allowing for land cover analysis, change detection, and environmental monitoring.

Beyond these core data types, geospatial data can also include:

- **Elevation Data:** Representing the height or depth of a location relative to a reference level (e.g., sea level). This is crucial for topographic maps, flood risk assessments, and 3D visualization.

Attribute Data: Descriptive information associated with geospatial features. For instance, a POI representing a heritage site or an infrastructure might have attributes like name, building technique types, and users. Geospatial data can be categorized into different types:

- **Vector Data**: Represents points (e.g., cities), lines (e.g., roads), and polygons (e.g., country boundaries).
- **Raster Data**: Consists of a grid of cells, where each cell contains a value representing a specific attribute (e.g., temperature, elevation).

3.4.1. GIS file formats

GIS file formats are specialised formats used to store and manage geospatial data (Bolstad, 2012). These formats are designed to capture, store, manipulate, and analyse geographic information efficiently. The most commonly used GIS file formats are:

- **Shapefile**: developed by Esri, shapefiles are one of the most widely used GIS formats. They consist of multiple files (.shp, .shx, .dbf) that store geometric data (points, lines, polygons) and attribute data (attributes associated with the geometries). Suitable for storing both vector and attribute data in a single file set.

- **GeoJSON**: open standard format based on JSON (JavaScript Object Notation). It supports various geometries (points, lines, polygons) and their properties (attributes). Widely used for web mapping applications due to its simplicity and compatibility with JavaScript.

- **Keyhole Markup Language**: Developed for Google Earth, KML is an XML-based format used to display geographic data in an Earth browser. KMZ is a compressed version of KML that includes all necessary files (KML, images, icons). Used for visualizing geographic data with 3D imagery, annotations, and overlays.

- GeoTIFF: GeoTIFF is a format that embeds geographic metadata within a TIFF (Tagged Image File Format) file. It supports raster data (gridded images) with spatial information such as coordinate system, extent, and projection. Commonly used for satellite imagery, aerial photography, and elevation models.

- **File Geodatabase**: Developed by Esri, File Geodatabase is a container format for storing multiple datasets (feature classes, tables, raster data) within a single file system folder. It provides efficient storage, data compression, and support for complex data relationships. Used in Esri's ArcGIS software for managing and organizing GIS data.

- **PostGIS**: PostGIS is a spatial database extension for PostgreSQL, allowing storage, manipulation, and analysis of spatial data. Supports both vector (points, lines, polygons) and raster data. Offers advanced spatial functions and capabilities for spatial queries and analysis.

- **GPKG** (GeoPackage): GeoPackage is an open standard format defined by the Open Geospatial Consortium (OGC). It is a SQLite-based format that stores both vector and raster data, along with metadata, in a single file. It is specifically designed for interoperability and portability across different GIS software and platforms.

These GIS file formats cater to different types of geospatial data and applications, offering flexibility, efficiency, and compatibility with various GIS software systems and tools. RESILIAGE tools will choose the right format depending on the specific requirements of the project, including data type, volume, interoperability needs, and software compatibility. The RESILIAGE Atlas design is ongoing (T3.3) and a description will be provided by D3.1 RESILIAGE Resource Ecosyste, First Release (M12) while the overall specification will be provided by D3.3 (M32)

The following subsections analyse in detail the major formats envisaged in the RESILIAGE project.

3.4.1.1. GeoJSON

GeoJSON, or Geographic JavaScript Object Notation, is an plain-text format designed for representing vector geometries, with or without non-spatial attributes, based on the JavaScript Object Notation, JSON (Butler, 2016). GeoJSON has become a very popular data format in many GIS technologies and services related to web mapping, as depicted in Figure 4.

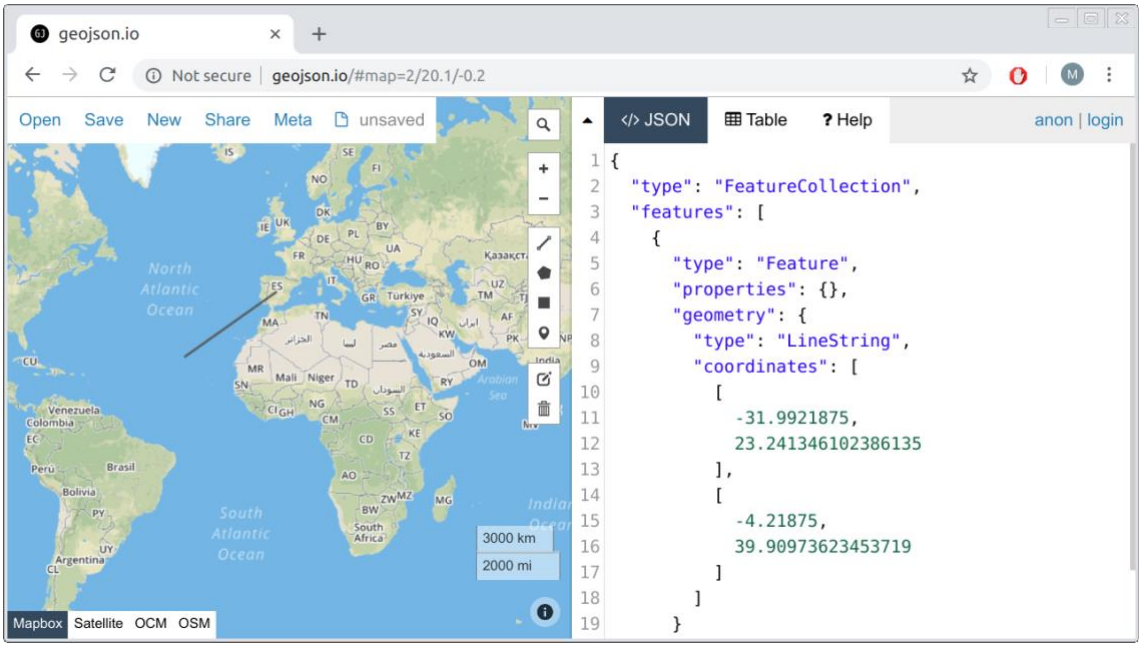

Figure 4 - Example of GeoJSON structure

It is actually the standard format for passing spatial vector layer data between the client and the server in web applications. This code is lightweight, easy to understand, and works seamlessly with various mapping tools and libraries.

GeoJSON's popularity for geospatial data stems from several key advantages. First, its simplicity makes it accessible. Because it leverages the familiar JSON format, even those without extensive coding knowledge can understand the structure of a GeoJSON file, making it easy to read and edit. Second, GeoJSON enjoys widespread support across various mapping platforms and libraries. As an open standard, you can share your geospatial data across different systems without compatibility issues. This universality is a major benefit for collaboration and data exchange. GeoJSON also

boasts flexibility. It can represent a wide range of geographic features, from points of interest like cafes and landmarks to lines such as roads and rivers, and even areas like city boundaries or national parks. This versatility makes it suitable for a broad spectrum of mapping applications. Finally, GeoJSON files are compact in size. This efficiency makes them ideal for storage and transmission over networks, especially when dealing with large datasets or real-time data streams.

3.4.1.2. Shapefile

A shapefile is a digital format for storing geographic information within a GIS. Developed by Esri, it is a widely used vector data format that allows for the storage of location, shape, and associated attributes of geographic features. Unlike raster data (like satellite images), which uses grids, vector data uses points, lines, and polygons to represent features (Rodrigue, 2024).

Introduced in the early 1990s, the shapefile format is one of the most common GIS vector data formats compatible with the majority of software platforms. It was designed as a compromise based on the most widely used database format of the time by indexing it with a feature file. Technically, a shapefile is as a collection of files, typically four, all working together. The main file, with a .ship extension, stores the actual geometric shapes of the features. An additional file, the .dbf file, holds the tabular attribute data, like the names or descriptions of features. There are also index files (.shx and .prj) that aid with data organization and reference coordinate systems, as depicted in Figure 5.

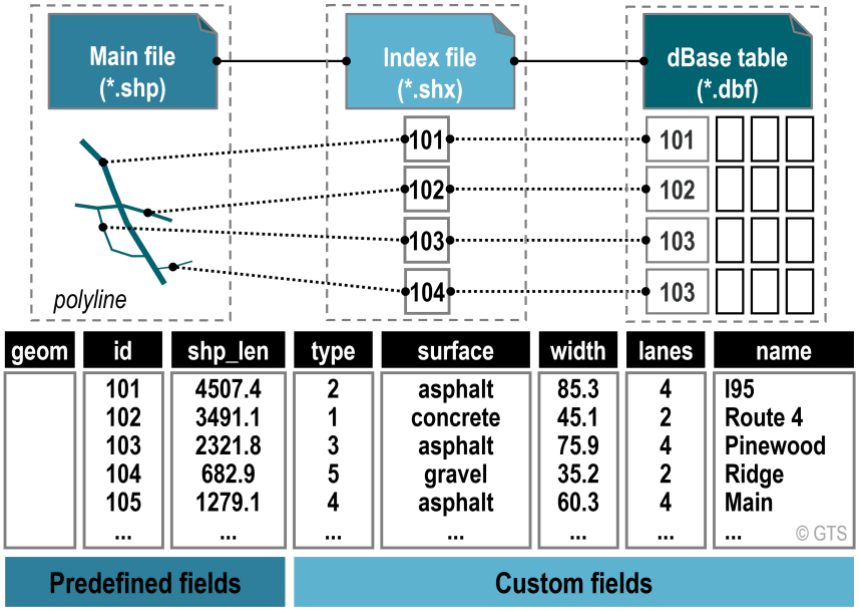

Figure 5 - The ESRI Shapefile Model

The strength of shapefiles lies in their simplicity and widespread compatibility. This openness allows them to be used across various GIS software programs, fostering data sharing and collaboration. Shapefiles are particularly useful for representing discrete features like buildings, roads, or environmental zones.

However, it's important to note that shapefiles have limitations. They cannot inherently store complex topological relationships between features, and file size can become an issue for very large datasets. Despite these limitations, shapefiles remain a popular and practical choice for storing and working with geographic vector data.

3.4.1.3. GeoTIFF

GeoTIFF, or Geospatial Tagged Image File Format, is a specialized file format designed for storing raster geospatial data. It builds upon the standard TIFF by embedding geographic metadata directly within the file. This metadata includes crucial information such as the coordinate system (projection), geographic extent (bounding box), pixel size (resolution), and sometimes georeferencing details (affine transformation parameters) (Ritter, 1997).

This format is highly versatile, accommodating various types of raster data such as satellite imagery, aerial photographs, digital elevation models (DEMs), and thematic maps. Its versatility makes it invaluable across disciplines like remote sensing, environmental monitoring, agriculture, and urban planning. One of GeoTIFF's strengths lies in its widespread support across GIS software and geospatial data processing tools. This ensures interoperability, allowing seamless exchange and use of geospatial data across different platforms and applications. GeoTIFF files also support compression methods like deflate and Lempel-Ziv-Welch, which help manage large datasets efficiently while maintaining data integrity. This capability is crucial for handling high-resolution imagery and other large-scale geospatial datasets.

Applications of GeoTIFF span diverse fields, including GIS mapping, terrain modeling, land cover classification, disaster management, and environmental monitoring. By embedding spatial metadata directly within the file, GeoTIFF facilitates accurate geospatial analysis and visualization, supporting informed decision-making and research in numerous domains.

3.4.1.4. GPKG

GeoPackage is an open standard file format developed by the Open Geospatial Consortium specifically for storing and sharing geospatial data. It leverages SQLite, a widely used embedded database engine, to accommodate various types of geographic information within a single, compact file (Rashidan, 2015).

Key features of GeoPackage include its support for both vector and raster data types. It can store points, lines, polygons, as well as gridded and tile pyramid raster data. This flexibility makes it suitable for storing diverse geospatial datasets such as maps, satellite imagery, and elevation models. GeoPackage includes built-in support for embedding spatial metadata. This metadata includes details like coordinate reference systems (CRS), spatial extents (bounding boxes), and data types, ensuring that the geospatial context of the data is preserved. Utilizing SQLite as its backend ensures GeoPackage files are cross-platform compatible and scalable. It supports standard SQL queries for data extraction and manipulation, making it accessible to a wide range of applications and workflows.

The format's self-contained nature simplifies data sharing and distribution, as all data and metadata reside within a single file. This portability is advantageous for field data collection, web mapping applications, and distributing standardized geospatial datasets

across organizations. As an OGC standard, GeoPackage promotes openness and interoperability in geospatial data exchange. It is widely adopted across the GIS community for its efficiency, versatility, and support for modern geospatial data management practices.

3.5. Data Lake technologies

RESILIAGE projects has been conceived with a centralized repository. A Data Lake is a centralised repository designed to store vast amounts of raw data in its native format. Unlike traditional data warehouses that store processed and structured data for specific uses, it retains all types of data – i.e., structured, semi-structured, and unstructured - in its original form until it's needed for analysis or other purposes (Miloslavskaya, 2016). The Figure 6 compares the major features of the major data storage systems (and paradigms).

	DATA WAREHOUSE	DATA LAKE
DATA TYPES	Structured, processed data from operational databases, applications and transactional systems	Structured, semistructured and unstructured data from sensors, apps, websites, etc.
PURPOSE	Predefined purpose for business intelligence, batch reporting and data visualization	May not have a predefined purpose; typically used for machine learning, deep analysis and discovery
USERS	Data engineers, business analysts, data analysts	Data engineers, data scientists
SCHEMA POSITION	Schema-on-write	Schema-on-read
BENEFITS	Categorized historical data stored in a single repository with ease of access for the end user	Data stored in its native format, allowing flexibility for data scientists to analyze and develop models from diverse data sources

Figure 6 - Data Lake vs Data Warehouse - Major differences

Key characteristics of a data lake include scalability, flexibility, and storage efficiency. Data lakes can scale horizontally, allowing organizations to manage massive data volumes from various sources without predefined schemas or fixed storage limits. They accommodate diverse data types such as structured data from databases, semi-structured data like JSON or XML, and unstructured data such as documents, images, and videos. Data lakes use cost-effective storage solutions like cloud storage (e.g., Amazon S3, Azure Blob Storage) or on-premises solutions (e.g., Hadoop Distributed File System, HDFS) to store large volumes of data economically. Unlike traditional databases that use schema-on-write (predefining data structure before storing), data lakes employ schema-on-read. This means data is stored as-is, and the schema is applied when the data is accessed or queried, providing flexibility in data analysis and exploration.

Components of a data lake include a storage layer (physical storage infrastructure), data ingestion processes, data processing and analytics frameworks (e.g., Apache Spark, Hadoop), and metadata management (catalogues and repositories for data lineage and usage information). Data lakes offer several benefits, including centralized data storage, flexibility for exploratory data analysis and machine learning, cost efficiency compared to traditional data warehouses, and scalability to accommodate growing data volumes and new data sources.

Use cases of data lakes include big data analytics, data science and machine learning initiatives, IoT and sensor data analysis, and enterprise data integration across business units and systems.

3.5.1. Data Lake Enabling Technologies

RESILIAGE centralised repository must be conceived to provide a consistent approach to complex information. For this purpose, it considers that a data lake isn't a single technology, but rather a strategic approach to data management (Giebler, 2019). It leverages a combination of powerful technologies to create a central repository for all your data, regardless of its format or origin. Here are some key technologies that make data lakes work:

- **Distributed Storage Systems:** Data lakes typically store massive datasets, often exceeding the capacity of traditional relational databases. Distributed storage systems like Hadoop Distributed File System (HDFS) provide a scalable and cost-effective solution for storing large volumes of data across multiple commodity servers.
- **Data Ingestion Tools:** Data from various sources, including databases, sensors, social media, and web logs, needs to be efficiently ingested into the data lake. Data ingestion tools provide mechanisms for extracting, transforming, and loading (ETL) data into the lake in a structured and organized manner. Popular examples include Apache Flume and Apache Kafka.
- **Data Management Frameworks:** Managing and organizing data within a data lake requires robust data management frameworks. These frameworks provide functionalities like schema management, data quality checks, and access control. Apache Hive and Apache Spark are widely used frameworks for data management in data lakes.
- **Data Processing Engines:** Data in its raw form within a data lake is rarely ready for immediate analysis. Data processing engines like Apache Spark and Apache Flink enable large-scale data processing, transformation, and manipulation to prepare the data for further analysis.
- **Metastore Services:** With vast amounts of data residing in the lake, keeping track of its location, format, and lineage becomes crucial. Metastore services like Apache Atlas act as a central registry, cataloging data assets within the data lake and providing valuable metadata for efficient data discovery and exploration.
- **Security Tools:** Security is paramount when dealing with sensitive data. Data lake architectures incorporate robust security tools to ensure data access control, encryption, and auditability. Techniques like role-based access control (RBAC) and data encryption at rest and in transit are essential for protecting data within the lake.

3.5.2. Big Data Storage systems

Traditional relational databases, while efficient for structured data, struggle with the sheer volume, variety, and velocity of big data. Big data storage systems address these challenges by offering:

- **Scalability:** The ability to grow seamlessly as data volumes increase.

- **Flexibility:** Accommodating diverse data formats, from structured tables to complex multimedia files.
- **Performance:** Enabling efficient data retrieval and processing for timely analytics.

Several big data storage systems cater to the specific needs of organizations. Here are some of the most prominent options:

- **Distributed File Systems (DFS):** These systems distribute data across multiple commodity servers, providing scalability and fault tolerance.
  - **Hadoop Distributed File System (HDFS):** A widely used open-source DFS known for its high throughput and scalability.
- **NoSQL Databases:** These offer a flexible alternative to relational databases, handling diverse data structures and scaling horizontally across multiple servers.
  - **Amazon DynamoDB:** A highly available and scalable NoSQL database service offered by AWS.
  - **MongoDB:** A popular open-source NoSQL document database known for its flexibility and ease of use.
- **Object Storage:** This approach stores data as self-contained objects with associated metadata, enabling efficient management of unstructured and semi-structured data.
  - **Amazon S3:** A widely adopted object storage service by AWS, offering scalability, durability, and cost-effectiveness for storing a variety of data.
  - **Google Cloud Storage:** A scalable and cost-efficient object storage service from Google Cloud Platform, ideal for storing large datasets and media archives.

3.5.3. Database with geospatial features

RESILIAGE especially refers to regions endangered by natural disasters and climate change related issues, including flooding, wildfires, earthquakes, landslide, heatwaves. Inherently spatial elements are relevant information. Locations, distances, and relationships between places hold valuable information across various sites. For these purposes, geospatial databases emerge as powerful tools for managing and analysing data with a geographic component.

Geospatial DBs, also known as spatial DBs or GIS DBs, are specialized DBs designed to store and manage data that has a geographic reference. This data can include:

- **Points of Interest (POI):** Locations of specific features like buildings, parks, or restaurants.
- **Lines:** Representing features like roads, rivers, or pipelines.
- **Polygons:** Defining areas like city boundaries, land parcels, or environmental zones.

Unlike traditional databases that focus on storing alphanumeric data, geospatial databases incorporate spatial data types like points (longitude/latitude coordinates), linestrings (sequences of points), and polygons (closed loops defining areas). This allows

them to not only store location information but also perform spatial queries and analyses based on geographic relationships.

The key features of Geospatial Databases are:

- **Spatial Data Types:** As mentioned earlier, geospatial databases support specialized data types to represent points, lines, and areas on a map.
- **Spatial Indexing:** They employ spatial indexing techniques to optimize queries involving location. This enables efficient retrieval of data based on geographic criteria (e.g., find all restaurants within a 5-kilometer radius).
- **Spatial Queries:** Geospatial databases allow you to perform complex queries based on spatial relationships. You can search for data within specific areas, find nearest neighbors, or analyze spatial patterns.
- **Integration with GIS Software:** These databases often integrate seamlessly with GIS software, allowing for visualization and analysis of spatial data on maps.

3.5.4. Data Catalogue solutions

Data Catalogues are essentially searchable inventories of organisation's data assets. They function like detailed library catalogues, providing users with information about the data's location, format, content, and lineage (origin and transformations). This metadata – "data about data" – is crucial for efficient data discovery, understanding, and utilization.

Benefits of Data Catalogues are the followings:

- **Improved Data Discovery:** Users can easily search and find relevant data assets, eliminating time wasted hunting for unknown datasets.
- **Enhanced Data Governance:** Data catalogues promote better data governance by facilitating data lineage tracking, access control, and data quality monitoring.
- **Increased Data Collaboration:** By fostering data transparency and discoverability, data catalogues encourage collaboration between teams and departments.
- **Boost Productivity:** Users spend less time searching and more time analyzing valuable data, leading to increased productivity.
- **Improved Data Quality:** Data catalogues can highlight inconsistencies or missing information within datasets, enabling proactive data quality management.

There are two main approaches to data catalog implementation:

- **Cloud-Based Data Catalogues:** These are managed services offered by cloud providers like AWS Lake Formation or Microsoft Azure Purview. They leverage automated crawlers to discover and ingest data from various cloud storage locations, automatically generating metadata.
- **Custom Data Catalogue Solutions:** Organizations can develop or deploy in-house data catalogue solutions. This approach offers greater flexibility and customization but requires more upfront investment and ongoing maintenance. One such custom solution (details remain confidential) can be tailored to integrate seamlessly with your specific data ecosystem and address unique data governance needs.

In the open-source software, the source code is released under a license in which the copyright holder grants users the rights to use, study, change, and distribute the software to anyone and for any purpose. On the contrary, proprietary software legally remains the property of the organization, group, or individual who created it, and the code is not publicly available. The following table (Table 1**Error! Reference source not found.**) compares open-source software with proprietary software according to the main aspects of security, stability, cost, warranty, community, and support.

Table 1 - Comparison between Open Source and Proprietary software

	Open-Source Software	Proprietary Software
Security	Because anyone can view and modify open-source software, any developer might spot and correct errors or omissions that a program's original authors might have missed.	Security is totally in charge of the creators.
Stability	Stable versions of the software are released but not at regular time intervals. In case creators stop working to the project, it continues to get implemented by the community	Creator gives software which it was probed. If the software house stops distributing the software it is no more available.
Warranty	Limited or no warranty	Full warranty
Community	Large community of developers around open-source software	None or small community
Cost	Code is available for free	High cost per license
Support	Support is given by the community of developers, but it is not assured	Support is given by the software house
Additional functionalities	It is possible to edit the code publicly available to add new functionalities	Beside the functionalities implemented by the authors, it is not possible to add new custom features.

3.6. Identification of the technical solutions

Due to the diverse nature of the data and the specific metadata requirements associated with various files, the chosen solution is to develop a custom implementation for the RESILIAGE Data Lake architecture. This bespoke solution is tailored to meet the unique needs of the project, ensuring efficient storage, retrieval, and management of both data and metadata.

The primary storage for the data is an AWS S3 bucket (Amazon Web Services, Amazon S3, 2024), a scalable and highly durable cloud storage service provided by Amazon Web Services. AWS S3 is well-suited for storing large volumes of diverse data types, including structured, semi-structured, and unstructured data. The flexibility of S3 allows it to

accommodate the varied datasets generated by different digital tools within the RESILIAGE project.

To manage the metadata, the solution employs a NoSQL database, specifically a MongoDB instance (MongoDB, 2024). MongoDB is renowned for its high performance, flexibility, and scalability, making it an ideal choice for this application. Unlike traditional relational databases that use tables and rows to store data, MongoDB employs a document-oriented data model. This model is particularly advantageous for applications where data structures can vary significantly and evolve over time. In MongoDB, data is stored in collections, which are groupings of documents. Each document is akin to a record in a relational database, but unlike relational tables, MongoDB collections do not enforce a fixed schema. This schema-less nature of MongoDB allows each document within a collection to have a different structure, making it highly adaptable to changing data requirements.

There are several benefits of using MongoDB for metadata storage, for example:

- **Flexible Schema**: MongoDB's schema-less nature provides significant flexibility in managing metadata. Traditional relational databases require a predefined schema, which can be limiting when dealing with varied and evolving data types. MongoDB collections can store documents with different structures, enabling a dynamic and adaptable data model. This flexibility is crucial for storing metadata that can vary greatly between different types of data and tools. As new tools are added or existing ones evolve, the metadata schema can easily be modified without requiring extensive database redesign.
- **Ease of Integration**: The ability to store diverse types of metadata without rigid schema constraints simplifies integration with various tools and data sources. This seamless integration facilitates a more adaptable cataloguing process, allowing the data lake to easily incorporate new data types and sources as they arise. By supporting a wide range of metadata structures, MongoDB enables efficient and streamlined data management, enhancing the overall functionality of the RESILIAGE data lake.
- **Scalability**: MongoDB is designed for horizontal scaling through a process known as sharding. Sharding distributes data across multiple servers, allowing the database to handle large volumes of data and high-throughput operations. This scalability is essential for a growing data lake, where the volume of metadata can increase rapidly. As the RESILIAGE project expands, MongoDB's ability to scale horizontally ensures that the metadata store can accommodate increasing data loads without compromising performance.

To maximize MongoDB's potential, the database will be organized into collections specific to each digital tool using the catalogue data. This approach leverages MongoDB's flexible schema and high performance, enabling each tool to store and query its specific metadata efficiently. For instance, the RESILIAGE project includes tools for environmental monitoring, disaster response, and preparedness planning. Each tool will have its dedicated collection within the MongoDB instance. This structure ensures that the catalogue remains organised and scalable, accommodating the diverse and dynamic nature of the metadata associated with different tools.

The custom-made solution for the RESILIAGE data lake architecture, combining AWS S3 for data storage and MongoDB for metadata management, offers a robust, scalable,

and flexible approach to handling the project's diverse data needs. This architecture not only supports efficient data and metadata management but also ensures seamless integration and scalability, vital for the project's success and growth.

The technical choices that have been taken in this Section could change in case newer and most performant technologies will emerge during the implementation activity.

4. The RESILIAGE Data Lake

In this Chapter, a detail description of the solution implemented for the Data Lake of RESILIAGE project is given.

4.1. Architecture

The RESILIAGE Data Lake is designed to serve as a repository for storing both general data and geospatial data generated within the project. Users can upload data to the platform via exposed APIs or a web-based interface, which provides functionalities for uploading, retrieving, and filtering data and metadata.

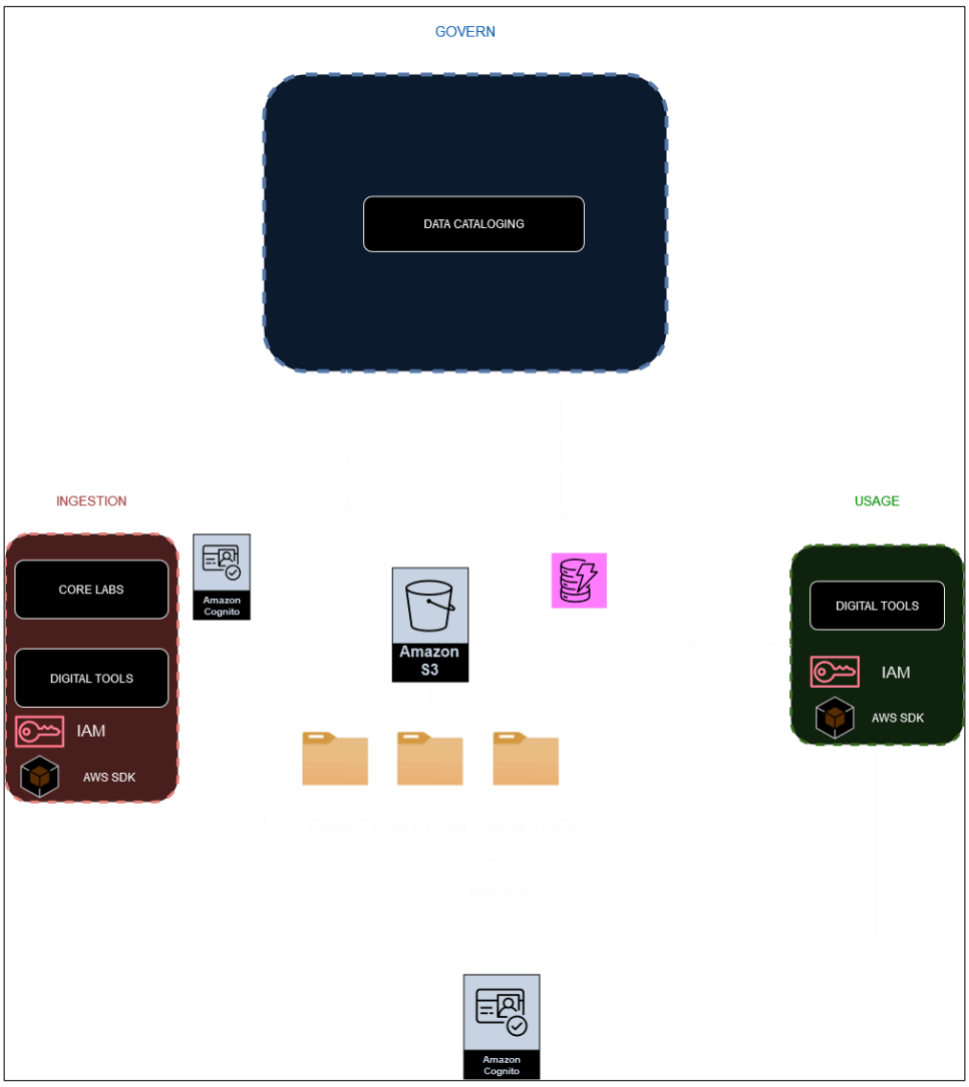

Figure 7 – RESILIAGE data lake general overview

Figure 7 – RESILIAGE data lake general overviewillustrates the architectural details. In detail, project data lake is composed by four main parts:

- **Ingestion:** The data will be ingested by users, agents running on digital tools ofthe RESILIAGE Resource Ecosystem such as the RAISE tool, the CORE Digital Network tool (T3.2, T3.3) and by the CORE labs;
- **Storage:** Where the data is allocated; an AWS S3 bucket (Amazon Web Services, Amazon S3, 2024); each type of data (based on what tool uses it) will be allocated in a specific S3 directory. The data cataloguing will be performed in a no-relational DB;
- **Govern:** Data must be organized so that it can be used by other tools without constant maintenance. This includes the process of cataloguing, that will be performed using the file's metadata;
- **Usage:** Where the data is used. In our case, by the tools of the whole RESILIAGE Resource ecosystem that includes several tools to be defined and specified by D3.1, D3.2 and D3.3 (e.g., the Monitoring Dynamic Resilience Dashboard, the DSS, Multidimensional Atlas for Communities Resilience, etc.). To retrieve this data 3 AWS technologies will be used: AWS Cognito (Amazon Web Services, Amazon Cognito, 2024), IAM permissions (Amazon Web Services, Policies and permissions in IAM, 2024) and AWS SDK (Amazon Web Services, Tools to Build on AWS, 2024). The first two are to check if a user and the tool has the required permissions to visualize or insert data, the last one will actively retrieve-provide this data.

In the ingestion phase, data enters the system through several channels. Users can manually upload data via web interfaces or APIs. Agents running on various tools also automatically ingest data into the system. Additionally, CORE labs, which are specialised units within the consortium, contribute data collected together with related research and operational activities. This multi-source ingestion approach ensures a diverse and continuous influx of data, enriching the RESILIAGE Data Lake with valuable and varied datasets.

For storage, the data is allocated in an AWS S3 bucket, a highly scalable and durable cloud storage service. Each type of data, depending on the tool that uses it, is organized into specific directories within the S3 bucket. For instance, data generated by different tools or for different purposes (such as logs, user data, or analytical results) will be stored in separate S3 directories to maintain organisation and ease of access. This structure allows for efficient data management and quick retrieval of relevant datasets when needed. The cataloguing of data is handled by a non-relational database, which provides the flexibility required to manage the diverse and often unstructured nature of Big Data. Non-relational DBs, such as Amazon DynamoDB or MongoDB, can store complex and varied data formats without requiring a fixed schema, making them ideal for dynamic and evolving data storage needs.

Governance involves organising and managing data so it can be used by other tools without constant maintenance. This includes the process of cataloguing, which is based on the metadata associated with each file. Metadata provides essential information about the data, such as its source, creation date, format, and any relevant tags or descriptions. Effective metadata management ensures that data is easily searchable and retrievable, facilitating efficient data use and integration across various applications. Governance also encompasses data quality control, compliance with data standards, and adherence

to security and privacy regulations. By implementing robust data governance practices, the Data Lake maintains high data quality and integrity, ensuring that the data remains reliable and useful for analytical and operational purposes.

In the usage phase, data stored in the Data Lake is utilised by various tools within the entire ecosystem. These tools may include data analytics platforms, machine learning models, business intelligence tools, and other applications that require access to large volumes of data. To retrieve and use this data, three key AWS technologies are employed: AWS Cognito, IAM (Identity and Access Management) permissions, and AWS SDK (Software Development Kit). AWS Cognito is used for user authentication and authorization, ensuring that only legitimate users can access the data. IAM permissions are set up to control access at a granular level, defining which users or tools have the right to view, insert, update, or delete specific datasets. This fine-grained access control is crucial for maintaining data security and preventing unauthorized access or modifications. The AWS SDK is a collection of tools and libraries that allow developers to interact programmatically with AWS services. It provides the functionality needed to retrieve and manipulate data stored in the S3 bucket, making it accessible to the various tools and applications within the ecosystem.

4.2. Metadata personalisation

A well-curated data source is fundamental for ensuring efficient data recovery and systematic organization, which are critical for maximizing the usability and value of the data. Proper curation involves not only collecting and storing data but also tagging and cataloguing it in a way that makes it easily accessible and understandable for end users. This process enhances the ability to quickly retrieve relevant data, supports accurate analysis, and facilitates effective decision-making. The data in such a repository can encompass a wide variety of elements, each serving different purposes and applications within the organization.

One category of data within this repository includes PDF documents. These documents can contain a wealth of information such as best practices, guidelines, research papers, reports, user manuals, and other documents that users might upload. By organizing these PDFs systematically, users can easily find the specific documents they need, whether they are looking for historical data, reference materials, or operational guidelines. Proper metadata tagging, such as titles, authors, publication dates, and relevant keywords, ensures that these documents are easily searchable and accessible.

Images and videos form another important category of data. These could include photographs, diagrams, charts, graphs, and informative videos. Images and videos are particularly useful for visualizing complex data, illustrating key concepts, or providing training and instructional content. For instance, charts and graphs can be used to represent statistical data visually, making it easier to interpret trends and patterns. Videos might include tutorials, presentations, or recorded webinars that provide in-depth explanations of various topics. Organizing these multimedia files with appropriate metadata such as descriptions, tags, and timestamps helps users quickly locate and utilize the visual content they need.

Meteorological data is also a valuable addition to a well-curated data source, especially for weather-related tools and applications. This data can include various weather parameters such as temperature, humidity, wind speed, precipitation, and atmospheric pressure. By incorporating meteorological data, users can add an insightful layer of information to their analyses and presentations. In disaster management, accurate weather data is crucial for predicting and preparing for extreme weather events. Properly organizing and labeling this data with fields such as location, date, and specific weather parameters ensures that it can be effectively used in relevant applications.

Numerical data, often contained in .csv (comma-separated values) files, is another crucial element. CSV files are widely used for storing tabular data in a simple text format, making them easy to read and process. This numerical data can encompass a wide range of information, such as financial records, sales data, experimental results, or survey responses. To ensure this data is useful, it must be organized and labeled with appropriate fields such as column headers, units of measurement, and relevant categories. Proper curation of numerical data enables efficient analysis, accurate reporting, and data-driven decision-making.

To make the tools effective, it is essential that all types of data within the repository are organized and labeled with the appropriate fields. This involves implementing a comprehensive data management strategy that includes metadata tagging, indexing, and cataloging. Each data element should have clear, descriptive metadata that explains its content, source, and relevance. For instance, PDFs should be tagged with relevant keywords, authorship details, and publication dates; images and videos should have descriptive tags, captions, and timestamps; meteorological data should include precise location and time information; and CSV files should have clearly defined headers and units of measurement.

Additionally, implementing a robust search and retrieval system within the data repository can significantly enhance data accessibility. Users should be able to perform keyword searches, filter data based on specific criteria, and retrieve the data they need quickly and efficiently. Advanced search functionalities, such as faceted search and full-text search, can further improve the user experience by allowing more refined and targeted searches.

4.3. User Authentication

In the realm of web and mobile application development, managing user authentication and access control can be a complex endeavour. Here's where AWS Cognito steps in, a service offered by AWS that simplifies this process.

Considered a Customer Identity and Access Management (CIAM) solution, Cognito offers a comprehensive suite of features to address user identity management within your applications. These functionalities encompass:

- **User Registration and Login:** Cognito provides a built-in user directory for storing user information and credentials. Users can register for your application directly through Cognito, creating a secure login process.

- **Social and Enterprise Authentication:** Beyond traditional registration, Cognito integrates with social identity providers like Google and Facebook, allowing users to sign in using their existing social media credentials. Additionally, it supports federation with enterprise directories like Active Directory, facilitating seamless login for users within your organization.
- **Password Management:** Cognito enforces strong password policies and offers features like password reset to ensure the security of user accounts.
- **Authorization and Access Control:** A crucial aspect of user management, Cognito allows you to define granular access controls. This enables you to determine which users or groups have access to specific resources within your application, ensuring data security and application integrity.
- **Scalability and Performance:** Built for robust user management, Cognito can handle millions of users without compromising on performance. This scalability is ideal for applications with large user bases.

4.4. Data Management System customisation

Data lakes are vast repositories that store large volumes of raw data in its native format until it is needed. Efficiently managing and organizing this data is crucial for maximizing its value. Customizing a Data Management System (DMS) in data lakes involves thinking and implementing various tools and techniques to catalogue, organize, and process data to make it easily accessible and useful. Metadata directly from backend-side can play a critical role in this customization process.

Cataloguing is a critical component of data management in data lakes. It involves creating metadata that describes the data's structure and attributes, making it easier to find and use. Effective cataloguing enables users to quickly locate and retrieve data, ensuring that they can leverage the full potential of their data lake.

When considering solutions for cataloguing a data lake, AWS Glue (Amazon Web Services, What is AWS Glue?, 2024) and custom metadata compiled by the backend represent two primary options. AWS Glue is a fully managed ETL service that includes a data catalogue component, capable of automatically discovering and cataloguing metadata as it crawls the data stored in the data lake. The primary advantage of AWS Glue lies in its automated discovery capabilities, which allow Glue crawlers to detect new data and update the catalogue without manual intervention. Moreover, AWS Glue integrates seamlessly with other AWS services, providing a central repository for metadata and reducing operational overhead through its managed infrastructure. However, AWS Glue requires an initial setup involving the configuration of crawlers, jobs, and databases, which can be complex and potentially costly, depending on usage patterns.

On the other hand, custom metadata compiled by the backend involves creating and managing metadata directly within backend systems, with the metadata stored in a non-relational database. This approach offers significant flexibility, allowing for highly customized metadata structures tailored to specific needs and use cases. It also provides full control over the metadata generation, storage, and retrieval processes, enabling tight integration with existing backend processes and systems. Despite these advantages, this method demands considerable development effort to design and implement an effective metadata management system. Furthermore, ongoing maintenance and

updates are necessary to ensure the system continues to meet requirements, and scalability can pose challenges, as handling large volumes of data and metadata may require additional resources and architectural considerations.

In conclusion, the technology that will be used for metadata storage a custom-made software that will save this data in the no-relational database – i.e., MongoDB.

Figure 8 visually represents the crucial process of data ingestion into the data lake, illustrating the systematic flow of how data files are brought into the centralized repository for storage and subsequent utilization.

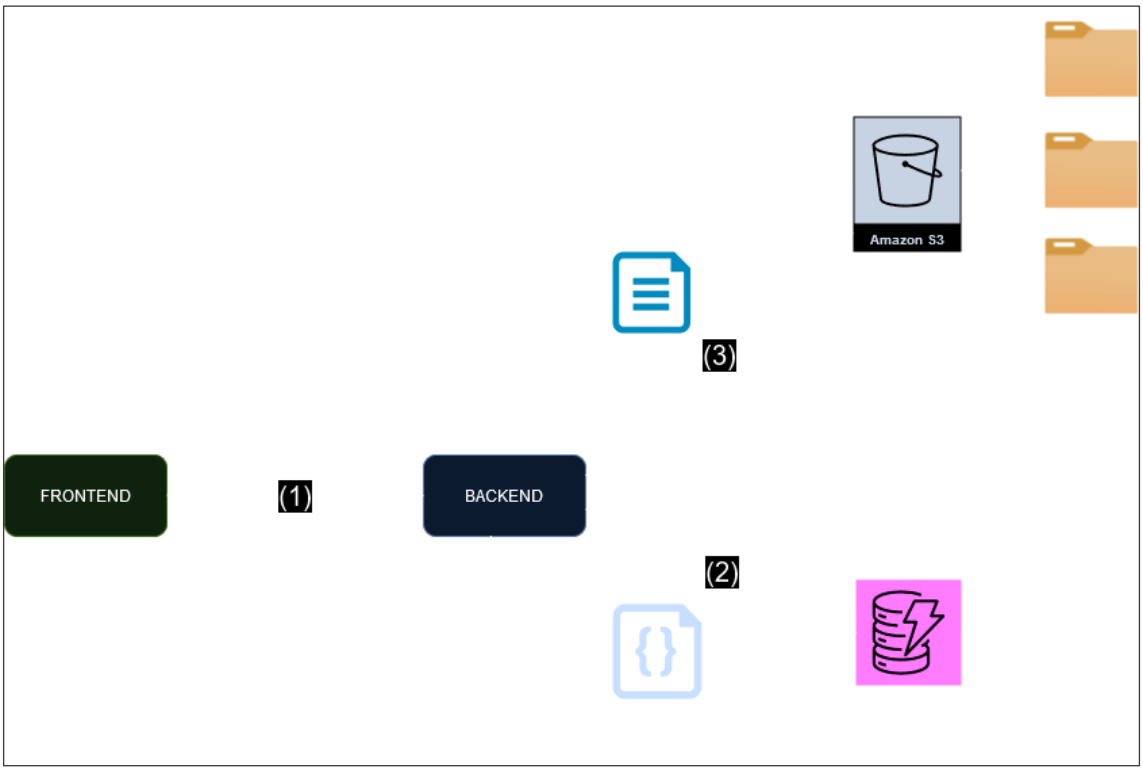

Figure 8 - Data lake file ingestion

This diagram provides a comprehensive overview of the various stages and components involved in the ingestion process, in detail:

- **(1)**: The user will make the request to put an object into the Data lake;
- **(1.1)**: The backend will check with Cognito if the user has such permissions;
- **(2-3)**: If the user has this privilege, the backend will generate a unique random key for the object. This key, along with other metadata values, will be stored in the non-relational database, while the object itself will be stored in the appropriate directory of S3;
- **(4)**: The software will give a positive or negative response based on the success of the operation.

Figure 9, indeed, illustrates a simplified version of how the catalogue will look like:

Figure 9 – Example of catalogue extract

With the data divided based on what tool uses it, and each loaded file with its unique ID stored along with other type of metadata (like the type of file). In essence, this catalogue structure serves as a foundational tool for navigating and leveraging the wealth of data stored within the organization's data lake, promoting informed decision-making and strategic insights across all levels of the organization.

4.4.1. User Interfaces

As mentioned before, the three main technologies that will display or enable the insertion of data to users are AWS Cognito, the AWS SDK, and AWS IAM permissions.

AWS Cognito will handle user authentication, ensuring that only authorized users can access the system. IAM permissions will define what these authenticated users are allowed to do, specifying the permitted actions on the S3 bucket, such as reading, writing

data, or both. The Amazon SDK is then used within the user interface to interact with S3, leveraging the temporary credentials provided by Cognito to securely perform the defined actions. This integration ensures that the system remains secure and user interactions with the data are efficiently managed, maintaining a high level of security and control over data access and manipulation. By using these technologies in conjunction, we can ensure that user authentication and authorisation are robust, and that data operations are conducted seamlessly within the specified security parameters.

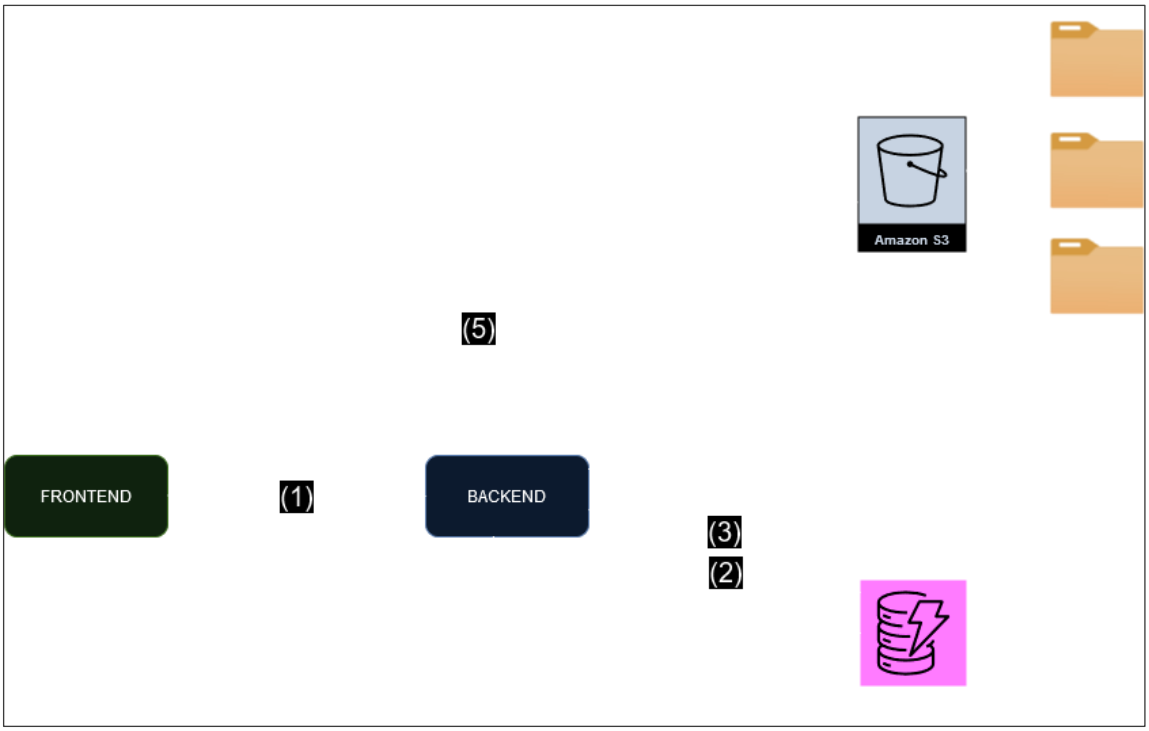

Figure 10 - Data lake file retrieval

Figure 10 depicts how the tools will generally access data from the Data Lake:

- **(1):** The user, through the frontend, performs a request to insert a file in the data lake. This request will be forwarded to the backend;
- **(2)**: The backend will use the parameters given by the frontend to search in the database for the index of the file in the S3;
- **(3)**: The database will return one or more items, depending on the original request
- **(4)**: The backend will use its IAM credentials to obtain a temporary pre-signed URL that will be given to the frontend to access the files in the S3. This practice will ensure a higher level of security for the system;
- **(5)**: The frontend will use this link to access the requested file(s). The link will expire after some time, depending on the software configurations.

4.5. Security

Security will be a fundamental aspect of the data lake implementation. The entire architecture will reside within its own Virtual Private Cloud (VPC), ensuring that resources within the VPC can communicate with each other while rejecting traffic from outside the network by being placed in private subnets.

Access to these services from external sources will be exclusively through an internet gateway positioned within a reverse proxy/load balancer, which will be the only public subnet. This reverse proxy will serve as the entry point from the outside to the rest of the platform, ensuring that all incoming traffic is properly routed and distributed to the appropriate backend services while providing enhanced security, scalability, and load management capabilities.

The connection from the VPC to the bucket will be securely established through AWS VPC endpoints, a service that functions similarly to a NAT gateway.

To ensure the highest level of security for file access in S3, all requests will be rigorously validated through IAM (Identity and Access Management) permissions. These permissions are configured so that only our proprietary software can access the S3 buckets, preventing unauthorized access. In addition to stringent access controls, the types of data permitted in the data lake are strictly regulated. This regulation is enforced by the backend systems of the ingestion tools, which are responsible for validating and sanitizing incoming data before storage. By implementing these robust security measures, we ensure that only approved data types are ingested, maintaining the integrity and security of our data lake. Finally, each piece of software will undergo thorough testing before its public release.

5. Conclusion

This deliverable aims to present the implemented solution for effective data management within the framework of the RESILIAGE project. The report detailed and demonstrated the deployed solution, accompanied by a comprehensive survey of cutting-edge technologies in data storage and management. By exploring and comparing these advanced technologies and approaches, the report provided insights into their capabilities and suitability for meeting the project's specific requirements.

Beginning with an extensive introduction to the essential components of the data management solution, particular emphasis was placed on the concept of a Data Lake. The report elucidated its structural framework, operational functionalities, and highlighted the significant benefits it offers. The comparison between in-cloud solutions and traditional on-premises data storage systems underscored the superior scalability, flexibility, and cost-effectiveness of cloud-based approaches as a critical advantage for handling the diverse and voluminous datasets central to the RESILIAGE project. Moreover, the report delved into the pivotal role of the data catalogue—a cornerstone for efficient data management. It examined various existing solutions, providing a comprehensive analysis of their features and functionalities. Emphasising their role in organising, discovering, and governing data, the discussion illustrated how these solutions contribute to optimising the search and retrieval processes within the RESILIAGE Data Lake environment. Central to this optimisation is the meticulous definition and implementation of metadata, ensuring that it encapsulates essential information and facilitates agile data querying.

Due to the diverse nature of the data and specific metadata requirements associated with various files, the chosen solution involved developing a custom implementation for the RESILIAGE data lake architecture. This bespoke solution is tailored to meet the unique needs of the project, ensuring efficient storage, retrieval, and management of both data and metadata. The primary storage solution for data within the RESILIAGE project is an AWS S3 bucket, leveraging its scalability and durability to accommodate large volumes of structured, semi-structured, and unstructured data generated by different digital tools. To manage metadata effectively, the solution integrates a MongoDB instance, renowned for its high performance, flexibility, and scalability. MongoDB's document-oriented data model allows for a flexible schema, facilitating dynamic adaptation to evolving data structures without the constraints of traditional relational databases. This flexibility supports efficient integration with diverse tools and data sources, enhancing the overall functionality of the RESILIAGE Data Lake.

MongoDB's scalability through horizontal scaling via sharding ensures the metadata store can handle increasing data volumes as the RESILIAGE project expands. Organising MongoDB into collections specific to each digital tool optimizes metadata storage and querying efficiency, ensuring the catalogue remains organized and scalable across different project domains.

In conclusion, this report represents a significant contribution to understanding the robust data management infrastructure implemented in the RESILIAGE project. It documents a suite of integrated components and methodologies designed to ensurehighly reliable data storage and efficient information retrieval. By adopting optimal solutions in data storage and management, the report provides comprehensive technical documentation of the implemented Data Lake and associated processes. This documentation enhances

comprehension and reinforces the data management solution's reliability and resilience, thereby advancing the overarching objectives of the RESILIAGE project.

It's important to note that the technical choices outlined in this report are based on current technologies and best practices available at the time of delivering. As the implementation progresses, consideration will be given to adopting newer and more performant technologies that may emerge, ensuring ongoing optimization and alignment with the project's evolving needs and goals.

Bibliography

Amazon Web Services, I. (2024). Retrieved from Amazon S3: https://aws.amazon.com/s3/

Amazon Web Services, I. (2024). Retrieved from Amazon Cognito: https://aws.amazon.com/cognito/?nc1=h_ls

Amazon Web Services, I. (2024). Retrieved from Policies and permissions in IAM: https://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies.html

Amazon Web Services, I. (2024). Retrieved from Tools to Build on AWS: https://aws.amazon.com/developer/tools/?nc1=h_ls

Amazon Web Services, I. (2024). Retrieved from What is AWS Glue?: https://docs.aws.amazon.com/glue/latest/dg/what-is-glue.html

Bolstad, P. (2012). *GIS fundamentals (Vol. 4).* White Bear Lake, MN: Eider Press.

Butler, H. D. (2016). *GeoJSON*. Retrieved from The GeoJSON Format: https://www.rfc-editor.org/rfc/rfc7946

DEPUTIES, M. (2017, 09 27). *Recommendation CM/Rec(2017)8 of the Committee of Ministers to member States on Big Data for culture, literacy and democracy*. Retrieved from CM/Rec(2017)8: https://search.coe.int/cm?i=0900001680750d68

Giebler, C. G. (2019). Leveraging the data lake: Current state and challenges. *Big Data Analytics and Knowledge Discovery: 21st International Conference, DaWaK 2019* (pp. 179-188). Linz, Austria: Springer International Publishing.

Guo, D. &. (2020). State-of-the-art geospatial information processing in NoSQL databases. *ISPRS International Journal of Geo-Information*, 331.

Miloslavskaya, N. &. (2016). Big data, fast data and data lake concepts. *Procedia Computer Science,*, 300-305.

MongoDB, I. (2024). Retrieved from MongoDB: https://www.mongodb.com/

Nugent-Folan, G. &. (2020). *Digitizing Cultural Complexity: Representing Rich Cultural Data in a Big Data Environment.*

Oliveira, B. D. (2024). Towards a Data Catalog for Data Analytics. *Procedia Computer Science*, 691-700.

Rashidan, M. H. (2015). GeoPackage as future ubiquitous GIS data format: a review. *Jurnal Teknologi*, 47-53.

Ritter, N. &. (1997). The GeoTiff data interchange standard for raster geographic images. *International Journal of Remote Sensing*, 1637-1647.

Rodrigue, D. J.-P. (2024). *The ESRI Shapefile Model*. Retrieved from transportgeography.org: https://transportgeography.org/contents/methods/network-data-models/esri-shapefile-model/

SAGIROGLU, S., & SINANC, D. (2013). Big data: A review. *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

Tsai, C. W. (2015). Big data analytics: a survey. *Journal of Big data*, 1-32.

Zeng, M. L. (2020). *Metadata.* American Library Association.

APPENDIX A – Data Lake Input Form

In the RESILIAGE Project, the management and organisation of datasets are paramount for ensuring the integrity, usability, and accessibility of data within the Data Lake. To achieve these goals, every dataset produced as part of the project must be accompanied by comprehensive metadata. This metadata serves as a critical tool for documenting the characteristics, origins, and structure of the datasets, facilitating efficient data management and retrieval.

To streamline this process and ensure compliance with the methodologies and paradigms established for the project, a specialized Data Lake input form has been developed. This form is designed to guide all partners responsible for producing or providing data, ensuring that they include all necessary metadata when submitting their datasets to the Data Lake.

The Table 2 details the structure of the metadata collection form designed and shared with partners.

Table 2 - Data Lake Input Form Structure

Data Lake Input Form	
Field of interest	**Description**
Dataset Name	The Name of the Dataset field provides a clear and concise title that uniquely identifies the dataset within the RESILIAGE Project. This name is essential for cataloguing, referencing, and retrieving the dataset in the Data Lake and other data management systems.
General Description of Dataset	The Brief Description of the Dataset field provides a succinct summary of the dataset's content, purpose, and key characteristics. This description gives potential users a quick overview of what the dataset includes and its relevance to their needs.
Purpose/Use Case	The Purpose or Intended Use of the Dataset field provides a detailed description of how the dataset is expected to be used within the context of the RESILIAGE Project. This information helps ensure that the dataset is applied appropriately and maximises its value to the project's goals and activities.

Source of Data		The Origin of the Dataset field provides detailed information about the source and provenance of the dataset. This includes specifying whether the dataset is newly collected, derived from pre-existing datasets, or obtained from third-party sources. Understanding the origin is crucial for assessing the dataset's credibility, relevance, and any potential limitations or biases.
Data Quality	**Data Quality Assessment: (is it ready to be used by a software?)**	The Data Quality Assessment field provides a detailed evaluation of the dataset's quality, determining its readiness for use by software applications. This assessment ensures that the data meets the necessary standards and is reliable. It also indicates whether the data can be used as is or whether pre-processing is required.
	Completeness: (e.g., are there any missing values?)	The Completeness field provides information about the dataset's integrity, specifically addressing whether all required data points are present and identifying any missing values. This assessment is crucial for understanding the dataset's reliability and suitability for various analyses.
	Accuracy: (e.g., known errors, validation checks)	The Accuracy field provides an evaluation of how closely the data values in the dataset match the true values or accepted standards. This field includes information about known errors, validation checks, and any measures taken to ensure data accuracy.
	Timeliness: (e.g., is the data up-to-date?)	The Timeliness field provides information about the currency and recency of the dataset. It assesses whether the data is up-to-date and available within the necessary timeframe for its intended use, ensuring its relevance and usefulness.
	Consistency: (e.g., are there any inconsistencies within the dataset?)	The Consistency field evaluates the uniformity and coherence of the dataset, focusing on identifying any discrepancies or inconsistencies within the data. It assesses whether the dataset adheres to predefined

		standards, formats, and rules, ensuring its reliability and usability for analytical purposes.
	Reliability: (e.g., is the source reliable and verified?)	The Reliability field assesses the trustworthiness and credibility of the dataset's source, ensuring that the data provided is accurate, consistent, and verified. It focuses on evaluating the reliability of the data collection methods, the reputation of the data sources, and any validation processes used to ensure data quality.
	Known Issues	The Known Issues field identifies and documents any known problems, limitations, or anomalies associated with the dataset. It provides transparency about potential issues that may affect the dataset's reliability, completeness, or accuracy.
	Data Cleaning Required: (Yes/No) - If yes, please describe.	The Data Cleaning Required field indicates whether the dataset requires preprocessing or data cleaning procedures to address inconsistencies, errors, or anomalies before it can be used effectively for analysis within the Resiliage Project.
File Format		The File Format field specifies the structure and type of the digital file containing the dataset. It defines how the data is organized, stored, and accessed, ensuring compatibility with software tools and facilitating data sharing and integration within the RESILIAGE Project.
Data structure information: (Table Names, Column Names and Data Types, Primary Keys…)		The Data Structure Information field provides a detailed description of the organization and schema of the dataset, including table names, column names, data types, and primary keys. This information outlines how the data is structured within the dataset, facilitating understanding and usage for data management and analysis.
Schema Information: (attach schema file if available)		The Schema Information field provides a structured description of the dataset's schema, detailing its tables, relationships, attributes (columns), data types, constraints, and other relevant

		metadata. This information is crucial for understanding how data is organized, stored, and related within the dataset, facilitating effective data management and analysis.
Data Volume	**Number of Records**	The Number of Records field provides the total count of individual data entries or rows within the dataset. It quantifies the volume of data available in the dataset, providing a basic measure of its scope and size.
	Data Size (approximate):	The Data Size (approximate) field provides an estimation of the total storage space occupied by the dataset. It quantifies the volume of data in terms of storage capacity, typically measured in megabytes (MB), gigabytes (GB), or terabytes (TB), depending on the dataset's size.
Keywords and Categorization	**Keywords: (relevant tags for searching and categorization)**	The Keywords field consists of relevant tags or terms used to categorize, search, and classify the dataset within the RESILIAGE Project. These keywords help describe the dataset's content, themes, and attributes, facilitating efficient data discovery, retrieval, and organization.
	Categories: (e.g., genre, type of content, topic)	The Categories field categorizes the dataset based on its genre, type of content, or overarching topic within the context of the RESILIAGE Project. It organizes the dataset into distinct classifications or domains, providing a structured framework for understanding its thematic focus and content.
Additional Notes/Comments		The Additional Notes/Comments field provides supplementary information, explanations, or context about the dataset that may not fit into other structured metadata fields. It allows for flexible documentation of important details, considerations, or caveats related to the dataset within the RESILIAGE Project.

By completing the Data Lake input form, partners ensure that their datasets are not only well-documented but also compatible with the overarching data management framework of the RESILIAGE Project. This meticulous approach to metadata inclusion promotes a

cohesive and efficient Data Lake environment, supporting robust data analytics and decision-making processes across the project.

APPENDIX B – Data Lake Input Examples

This appendix provides some concrete examples of the data collection processes description and methodologies employed for WP1 and WP2 within the RESILIAGE Project.

The main objective was to ensure comprehensive, accurate, and well-documented data inputs into the Data Lake, a centralized repository designed to support robust data management, analysis, and collaboration.

In the following sections, we present examples of how data from WP1 and WP2 were collected, annotated, and input into the Data Lake. These examples illustrate the practical application of our data collection framework, highlighting key metadata fields. Through these examples, we aim to provide clarity on the process and encourage adherence to these standards across all project partners, thereby ensuring the integrity and usability of the data within the Data Lake.

Here are the details of the data collected in WP1 and WP2,

Data collected within WP1:
T1.1 - International Disaster Risk Management (DRM) policies; Previous EU projects related to RESILIAGE
T1.2 - Scoping literature review; Lessons learned extracted from literature (LLs)

Data Collected within WP2:
T2.1 – Brainstorming Workshop Conceptualization of Community Resilience (First Responders) Data; Brainstorming Workshop Conceptualization of Community Resilience (Citizens Associations) Data; Brainstorming Workshop on Gender indicators Data; Workshop on Gender Multiscalar characterization in Disaster Risk Management (DRM) Data; Workshop on Gender Multiscalar characterization in DRR Data text transcriptions Data
T2.2 – Contact data of FG participants (Attendance Sheet, Registration Form); Policy template filled in by CORE partners; FG recordings; Interaction Map created in the FGs, FG protocols, Photos taken of the FGs
T2.3 – Cross sectional survey responses; Eye tracking experiment data
T2.4 – Virtual Reality (VR) experiment data; (Past Crisis Survey)
T2.5 – N/A
T2.6 – CORE Collaborative Workshop Local Knowledge Data; CORE Collaborative Workshop LLs extraction Data, CORE Collaborative Workshop Co-mapping Data Collaborative Workshop video recordings; CORE Collaborative Workshop recordings; Photos of participants taken in the Collaborative Workshops.

❖ WP1 - Data Lake Input Form Examples

Example of Data Lake Input Form Completion within WP1. This is a concrete example of the documentation supporting the collection and creation of datasets.

Dataset Name: International DRM policies
General Description of Dataset: Excel table listing key international disaster management policies by supra-national bodies and organisations
Purpose/Use Case: International policy framework as a baseline to analyse national policy development
Source of Data: VIC/POLITO/DEMIR
Data Quality:
- Data Quality Assessment: ready to be used by a software
- Completeness: focused on key policy makers and disaster management and climate change
- Accuracy: yes
- Timeliness: yes
- Consistency: yes
- Reliability: yes
- Known Issues: Depending on specific aspects of the DMC further policy areas could be added
- Data Cleaning Required: no

File Format: Excel
Data structure information:

Level	Type	Binding?	Chief audience/s	Link/s	Added by (Partner)

Schema Information: ?
Data Volume:
- Number of Records: 1 table with 106 entries
- Data Size (approximate): 50 KB

Keywords and Categorization:
- Keywords: DRM, international, policy
- Categories: DRM, policies, guidelines

Additional Notes/Comments:

Dataset Name: Previous EU projects related to resiliage
General Description of Dataset: Excel table listing all EU funded projects of the DRS calls since 2010, ranked via Delphi method by partners for Resiliage relevance, and accessible & relevant Resources listed by Resiliage topic
Purpose/Use Case: SOTA of previous EU projects relevant for any further project development to contrast what has been already done
Source of Data: DBL
Data Quality:
- Data Quality Assessment: ready to be used by a software
- Completeness: selection by relevancy to RESILIAGE; not all deliverables and project were publicly available
- Accuracy: yes, for what has been available
- Timeliness: yes
- Consistency: yes
- Reliability: yes, for what has been available
- Known Issues: unavailable projects and resources bias the sample, new projects and resources are continuing to emerge
- Data Cleaning Required: ?

File Format: Excel

Data structure information:

Human factors response / psychological aspects	"Crisis management, crisis communication"	Legal frameworks, policies	"Tech projects & Digital Solutions"	Policy recommendations, Communication guidelines	Soft solutions	Links to heritage	Links to climate change	Methodologies	Recommendations, lessons learnt, best practices

Schema Information: ?
Data Volume:
- Number of Records: 4 table with up to 160 single entries, presented in various ways
- Data Size (approximate): 100 KB

Keywords and Categorization:
- Keywords: DRS, previous, EU, projects
- Categories: DRS, previous, EU, projects

Additional Notes/Comments:

➢ WP2 - Data Lake Input Form Examples

Example of Data Lake Input Form Completion within WP2. This is a concrete example of the documentation supporting the collection and creation of datasets.

Dataset Name: Brainstorming Workshop Conceptualization of Community Resilience (FRs) Data
General Description of Dataset: Digital downloaded files of MIRO boards (software Miro) reporting portions of the digital boards results of workshop with FRs in consortium.
Source of Data: POLITO
Data Quality:
- Data Quality Assessment: ready to be transcribed into text format (csv or .rft / .doc).
- Completeness: all exercised fully completed by participants
- Accuracy: yes
- Timeliness: yes
- Consistency: yes
- Reliability: yes
- Known Issues:
- Data Cleaning Required: no

File Format: PDF
Data structure information: N/A
Schema Information:
Data Volume:
- Number of Records: approx. 10 pdf files for each digital Milro board
- Data Size (approximate): approx. 1 MB per pdf file

Keywords and Categorization:
- Keywords: Community Resilience, First Responders, DRR, risk scenarios, Cultural Heritage
- Categories: Conceptualization Community Resilience, Cultural Heritage

Additional Notes/Comments:

Dataset Name: Brainstorming Workshop Conceptualization of Community Resilience (CAs) Data
General Description of Dataset: Digital downloaded files of MIRO boards (software Miro) reporting portions of the digital boards results of workshop with citizen associations in consortium.
Source of Data: POLITO
Data Quality:
- Data Quality Assessment: ready to be transcribed into text format (csv or .rtf / .doc).
- Completeness: all exercised fully completed by participants
- Accuracy: yes
- Timeliness: yes
- Consistency: yes
- Reliability: yes
- Known Issues:
- Data Cleaning Required: no

File Format: PDF
Data structure information: N/A
Schema Information:
Data Volume:
- Number of Records: approx. 10 pdf files for each digital iIro board
- Data Size (approximate): approx. 1 MB per pdf file

Keywords and Categorization:
- Keywords: Community Resilience, Citizens Associations, DRR, risk scenarios, Cultural Heritage
- Categories: Conceptualization Community Resilience

Additional Notes/Comments:

Dataset Name: Brainstorming Workshop on Gender indicators Data
General Description of Dataset: Digital downloaded files of MIRO boards (software Miro) reporting portions of the digital boards results of workshop with task leaders.
Source of Data: POLITO
Data Quality:
- Data Quality Assessment: ready to be transcribed into text format (csv or .rft / .doc).
- Completeness: all exercised fully completed by participants
- Accuracy: yes
- Timeliness: yes
- Consistency: yes
- Reliability: yes
- Known Issues:
- Data Cleaning Required: no

File Format: PDF
Data structure information: N/A
Schema Information:
Data Volume:
- Number of Records: approx. 10 pdf files for each digital iIro board
- Data Size (approximate): approx. 1 MB per pdf file

Keywords and Categorization:
- Keywords: Community Resilience, Gender diversity, DRR, Cultural Heritage
- Categories: Conceptualization Community Resilience, Gender Diversity

Additional Notes/Comments:

Dataset Name: Workshop on Gender Multiscalar characterization in DRR Data
General Description of Dataset: Digital downloaded files of printed boards with results from the workshop on Gender in DRR with all consortium participants (in presence GA in Crete)
Source of Data: POLITO
Data Quality:
- Data Quality Assessment: ready to be transcribed into text format (csv or .rtf / .doc).
- Completeness: all exercised fully completed by participants
- Accuracy: some writings non fully readable
- Timeliness: yes
- Consistency: yes
- Reliability: yes
- Known Issues:
- Data Cleaning Required: no

File Format: JPG
Data structure information: N/A
Schema Information:
Data Volume:
- Number of Records: approx. 10 for each printed board
- Data Size (approximate): approx. 1 MB per jpg file

Keywords and Categorization:
- Keywords: Community Resilience, Gender diversity, DRR, Cultural Heritage
- Categories: Conceptualization Community Resilience, Gender Diversity

Additional Notes/Comments:

Dataset Name: Workshop on Gender Multiscalar characterization in DRR Data transcriptions Data
General Description of Dataset: Digital text file containing transcriptions of results of GA Workshop on Gender Multiscalar characterization in DRR.
Source of Data: POLITO
Data Quality:
- Data Quality Assessment:
- Completeness:
- Accuracy: some parts are not fully readable
- Timeliness: yes
- Consistency: yes
- Reliability: yes
- Known Issues:
- Data Cleaning Required: no

File Format: .doc / .pdf
Data structure information: N/A
Schema Information:
Data Volume:
- Number of Records: 1 doc 1 pdf
- Data Size (approximate): approx. 1 MB

Keywords and Categorization:
- Keywords: Community Resilience, Gender diversity, DRR, Cultural Heritage

- Categories: Conceptualization Community Resilience, Gender Diversity

Additional Notes/Comments:

Dataset Name: Contact Data of FG participants
General Description of Dataset: Excel table listing the names, contact details, Organisational affiliation & role, day of attendance of Field Study Focus Group participants per each CORE
Purpose/Use Case: Initial stakeholder list of each CORE & relevant organisations in crisis management who took part in the FGs
Source of Data: CORE lab responsible
Data Quality:
- Data Quality Assessment: ready to be used by a software
- Completeness: people who did not register to the event or could not attend the event but are relevant organisatiosn to crisis management are missing
- Accuracy: for the people who registered and came the information should be accurate
- Timeliness: it is documenting the participation pre-implementation of the FGs and could be updated for actual attendance on the date of the FGs (see Attendance Sheet; Registration Form)
- Consistency: yes
- Reliability: yes
- Known Issues: It is personal data required for implementing the FGs and further the communication and buildup of the CORE labs and limited to this purpose
- Data Cleaning Required: Yes, consolidate Contact data file with Attendance sheet and Registration form

File Format: Excel
Data structure information: Name, E-Mail, Affiliation, Attendance per day
Schema Information: ?
Data Volume:
- Number of Records: 1 table for each CORE listing between 15-40 entries
- Data Size (approximate): 50 KB

Keywords and Categorization:
- Keywords: CORE, lab, Field, Study, participant
- Categories: CORE lab participants, Field Study participants, Local Authorities, First line practitioners, Citizens

Additional Notes/Comments:
IT IS PERSONAL DATA PLEASE CHECK WITH D6.3 DATA MANAGEMENT THROUGHOUT THE DATA LIFE CYCLE BEFORE FURTHER PROCESSING THIS DATA OR PUBLISHING IT.

Dataset Name: **Policy template filled in by CORE partners**
General Description of Dataset: Excel table listing the national and local DRM policies and guidelines
Purpose/Use Case: Initial policy list of each CORE & relevant organisations in crisis management who took part in the FGs
Source of Data: CORE lab responsible
Data Quality:
- Data Quality Assessment: ready to be used by a software

- Completeness: local policy collection will be completed within T6.1 regarding all relevant polcies related to crisis management throught all DMC stages
- Accuracy: yes
- Timeliness: policy updates are need to be monitored
- Consistency: yes
- Reliability: yes
- Known Issues: N/A
- Data Cleaning Required: Yes, entering practices by COREs differ from each other (title, link, body, etc.)

File Format: Excel
Data structure information:

National/regional/local policy & regulation	Organisational guideline & protocols	Evaluation: best practices, lessons learned, recommendations
Name, Issuing agency, short description, reference link	Name, Issuing agency, date, short description, reference link	Name, Issuing agency, short description, reference link

Schema Information: N.A.
Data Volume:
- Number of Records: 1 table for each CORE listing between 10-40 entries
- Data Size (approximate): 50 KB

Keywords and Categorization:
- Keywords: CORE, lab, Local, National, Policy
- Categories: CORE lab, DRM policies

Additional Notes/Comments: N/A

Dataset Name: **Focus Grous (FG) recordings**
General Description of Dataset: Audio files recording the Focus Group sessions (usually 2 groups having 2 sessions of each 90 min)
Purpose/Use Case: FG discussion based on the Facilitators Guide serving T2.2 and T4.1
Source of Data: VIC/DBL
Data Quality:
- Data Quality Assessment: some parts unintelligble
- Completeness: yes
- Accuracy: yes
- Timeliness: yes
- Consistency: yes
- Reliability: yes
- Known Issues: It is personal data referencing direct and strong direct indicators, DO NOT PROCESS FURTHER.
- Data Cleaning Required: N/A

File Format: WAV
Data structure information: N/A
Schema Information: N.A.
Data Volume:
- Number of Records: 4 recordings per Field Study, occasionally up to 4 back up recordings with other devises
- Data Size (approximate): approx. 1GB per session recording

Keywords and Categorization:
- Keywords: CORE, lab, Focus Group, audio, recording
- Categories: CORE lab, Field Study, Focus Group

Additional Notes/Comments:
It contains personal data. To be checked and aligned with D8.3 data management throughout the data life cycle.

Dataset Name: **Interaction Map created in the Focus Groups (FGs) sessions**
General Description of Dataset: Photograph of an A0 sheet where focus group participants put their organisations onto and indicating their communication pathways and channels
Purpose/Use Case: FG discussion based on the Facilitators Guide serving T2.2 and T4.1
Source of Data: VIC/DBL
Data Quality:
- Data Quality Assessment: some groups misunderstood the exercise
- Completeness: some groups did not put all the actors on it
- Accuracy: some groups did not put all communication channels on it
- Timeliness: N/A
- Consistency: N/A
- Reliability: N/A
- Known Issues: it also contains personal data. It is a representation of their impression of the interactions, not of all interactions that do occur.
- Data Cleaning Required: N/A

File Format: JPG
Data structure information: N/A
Schema Information: ?
Data Volume:
- Number of Records: up to 3 interaction maps per group, 2 groups per field study, 5 field studies
- Data Size (approximate): approx.. 1,5 MB per photo

Keywords and Categorization:
- Keywords: CORE, lab, Focus Group, photo, Interaction map
- Categories: CORE lab, Field Study, Focus Group

Additional Notes/Comments:
It contains personal data To be checked and aligned with D8.3 data management throughout the data life cycle.

Dataset Name: Focus Group (FG) protocols
General Description of Dataset: Word files paraphrasing and quoting and summarising the Focus Group sessions (usually 2 groups having 2 sessions of each 90 min)
Purpose/Use Case: FG discussion based on the Facilitators Guide serving T2.2 and T4.1
Source of Data: VIC/DBL
Data Quality:
- Data Quality Assessment: some parts unintelligible
- Completeness: not all sessions have been fully transcribed
- Accuracy: paraphrasing and summarizing include an interpretation of the transcriber as to what the meaning and importance of a set of statements were
- Timeliness: yes
- Consistency: varying practices across Field Study sites

- Reliability: adequate
- Known Issues: Beyond the list of participants the protocols have not yet been systematically screened for any direct or strong indirect personal identifier to be removed.
- Data Cleaning Required: yes

File Format: DOCX
Data structure information: N/A
Schema Information: ?
Data Volume:
- Number of Records: 1 protocol per recording (4 recordings per Field Study, occasionally up to 4 back up recordings with other devises, 5 field studies)
- Data Size (approximate): approx. 100 KB per session recording

Keywords and Categorization:
- Keywords: CORE, lab, Focus Group, protocol
- Categories: CORE lab, Field Study, Focus Group

Additional Notes/Comments:
It can contain personal data To be checked and aligned with D8.3 data management throughout the data life cycle.

Dataset Name: **Photos taken in the FGs**
General Description of Dataset: Photograph of focus group participants during the focus group sessions
Purpose/Use Case: FG discussion based on the Facilitators Guide serving T2.2 and T4.1
Source of Data: VIC/DBL
Data Quality:
- Data Quality Assessment: N/A
- Completeness: There was no data collection aim, but for mere demonstrative purposes documenting the setting of the session and the interactions of the participants.
- Accuracy: N/A
- Timeliness: N/A
- Consistency: N/A
- Reliability: N/A
- Known Issues: it also contains personal data, however participants agreed to be photographed during the session. However, publishing the photos together with the FG protocols can enhance de-anonymisation.
- Data Cleaning Required: yes

File Format: JPG
Data structure information: N/A
Schema Information: ?
Data Volume:
- Number of Records: approx. 100-200 per field study
- Data Size (approximate): approx.. 1,5 MB per photo

Keywords and Categorization:
- Keywords: CORE, lab, Focus Group, photo
- Categories: CORE lab, Field Study, Focus Group

Additional Notes/Comments:
It can contain personal data To be checked and aligned with D8.3 data management throughout the data life cycle.

RESILIAGE
Revolutionising community resilience

OUR CONSORTIUM

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. Neither the European Union nor the granting authority can be held responsible for them.