Ref. Ares(2025)6076715 - 28/07/2025

DELIVERABLE 1.5: ANALYSIS OF UP-TO-DATE SYSTEMS OF PRIVATE DATA SHARING

WP1: Stakeholder Involvement in Data-Driven Analysis of the Food System in the Search for New Solutions

Responsible Author: EMY ZECCA — GINEVRA COLETTI (UNIFE)

Document information

Project number:	101134894
Project name:	Sustainability Optimization for Secure Food Systems
Project acronym:	SOSFood
Call:	Call: HORIZON-CL6-2023-GOVERNANCE-01
Type of action:	HORIZON-RIA
Service:	REA/B/04
Project starting date:	Fixed date: 1 February 2024
Project duration:	48 months
Project website:	https://www.sosfood-project.eu/
EU Project Officer:	Javier Martin-Membiela
Project Coordinator:	UNIVERSIDADE DE VIGO (UVIGO) LG CAMPUS LAGOAS
Address:	MARCOSENDE, VIGO PONTEVEDRA, 36310, SPAIN
Contact | Email:	jsimal@uvigo.es, eloy.miranda@uvigo.gal
Work Package:	Stakeholder Involvement in Data-Driven Analysis of the Food System in the Search for New Solutions
WP Lead Beneficiary:	Eurocoop
Relevant Task(s):	Task 1.3
Relevant Deliverable:	Deliverable 1.5
Relevant Milestone:	M6 – M48
Deliverable Lead Beneficiary:	University of Ferrara
Responsible Author:	Emy Zecca – Ginevra Coletti – Inga Vyšniauskienė

Document History

Time	Session	Date	Date
01	**Version 1**	**08/01/2025**	Emy Zecca – Ginevra Coletti
02	Version 2	16/06/2025	Emy Zecca – Ginevra Coletti
03	Version 3	27/06/2025	Emy Zecca – Ginevra Coletti– Inga Vyšniauskienė
04	Versione 4	08/07/2025	Emy Zecca – Ginevra Coletti– Inga Vyšniauskienė

Participants

Participants	Contact
Emy Zecca	emy.zecca@unife.it
Ginevra Coletti	ginevra.coletti@unife.it

Table of Contents

Executive Summary

1. Introduction 9

2. Methodology and matrix construction 11

3. Data sharing approaches 11

3.1 API (Application Programming Interface) 11

3.2 Data repository 13

3.3 Data Marketplace 15

3.4 File sharing services 16

3.5 Peer 2 Peer (P2P) Network 18

3.6 Data Collaboration Platform 19

3.7 Data federations and virtual databases 20

4. Comparison between different approaches 21

5. Conclusions 24

References 27

Executive Summary

In an era defined by pervasive digitalization, the strategic importance of data sharing has grown substantially across a diverse range of domains, including scientific research, healthcare, industry, and public governance. Data proliferates in both volume and strategic importance, organizations are increasingly required to adopt frameworks that facilitate secure, efficient, and collaborative modes of data exchange. This report presents a comprehensive analysis of the two principal paradigms that currently underpin data-sharing practices: data collaboration platforms and data federation and virtualization systems.

In the context of an increasingly digitized society, the strategic significance of data sharing has expanded markedly across multiple domains, including scientific research, healthcare, corporate enterprise, and governmental operations. As the volume and strategic value of data continue to grow, organizations are compelled to implement systems that enable secure, efficient, and collaborative mechanisms for data exchange. This report offers a thorough examination of the two predominant paradigms currently shaping data-sharing practices: data collaboration platforms and data federation/virtualization systems.

Data collaboration platforms represent centralized infrastructures designed to support real-time, multi-user engagement with shared datasets. These platforms are distinguished by their capacity to enable collaborative workflows through features such as role-based access control, version management, and integration with analytical and visualization tools. They promote transparency, accountability, and traceability, rendering them particularly effective in interdisciplinary and coordination-intensive environments. Nonetheless, their implementation entails specific challenges, notably concerning data centralization, compliance with privacy regulations, and the organizational changes required to support effective adoption.

In contrast, data federation and virtualization systems employ a decentralized approach that enables access to and querying of data residing in heterogeneous and distributed repositories, without necessitating physical relocation or duplication. By preserving data in situ and leveraging virtual integration layers, these systems reduce operational disruption and maintain existing governance frameworks. However, their deployment is associated with technical complexities such as schema heterogeneity, data synchronization difficulties, and regulatory constraints, especially in transnational data-sharing contexts.

Acknowledging that these two approaches are not inherently mutually exclusive, this report advocates for hybrid models that synthesize the collaborative advantages of centralized platforms with the flexibility and autonomy provided by federated systems. Organizations that adopt such integrated strategies—underpinned by strong governance

structures and comprehensive user training—are likely to be better positioned to transform data into actionable insights and sustainable competitive advantage.

To facilitate a structured comparison of these methodologies, the report introduces an analytical matrix comprising six evaluative dimensions: definitional framework, typology, operational mechanisms, core features, associated benefits, and implementation challenges. This framework enables a nuanced assessment of each data-sharing model and highlights the contexts in which each approach is optimally applied.

By aligning technical infrastructure with organizational objectives and regulatory obligations, this report aims to inform strategic decision-making and foster the development of intelligent, interoperable, and secure data ecosystems.

1. Introduction

In a world undergoing a digital transformation, where information exchange has become essential, data sharing stands as a foundational pillar that underpins numerous technological and organizational advancements. The ability to efficiently and securely disseminate data is not merely anymore just a technical concern, but a strategic imperative across diverse domains ranging from scientific research and healthcare to corporate enterprise systems and governmental operations (Kowalczyk and Shankar, 2011; Michener, 2015; Leonelli, 2016; Wickramasinghe & Bodendorf, 2019). As data continues to grow in both volume and strategic importance, the mechanisms and tools through which it is exchanged inevitably influence the effectiveness of collaborative endeavours, the robustness of data security protocols, and the overall pace of innovation (Kitchin, 2014; Borgman, 2015).

Economic actors face the challenge of not only accessing large volumes of information but also making that information usable, actionable, and secure across distributed teams and systems (Loshin, 2011; Mbuagbaw et al., 2017). Within this context, two main approaches have emerged: data collaboration platforms, and systems based on data federation and virtualization. Although both serve the broader purpose of enhancing data accessibility and usability, they can differ significantly in their conceptual framework, operational mechanisms, and areas of application. Understanding these differences is essential for organizations seeking to design or refine their data-sharing strategies in a way that aligns with evolving technological demands and regulatory expectations (Halevy et al., 2006; Parashar et al., 2020).

Data collaboration platforms are designed primarily to enable multiple users to interact with shared datasets in real time. These platforms aim is to centralize data within a common environment and provide tools that facilitate simultaneous editing, version tracking, and secure access control. Their most distinctive feature lies in their ability to support collaborative workflows, where multiple team members—often from different

departments or disciplines—can contribute to data analysis, transformation, and reporting. Real-time co-authoring, role-based permissions, and built-in audit trails foster a high level of transparency, accountability, and traceability. This makes them particularly well-suited for environments where collaborative decision-making is essential, such as in research institutions, business intelligence operations, and multidisciplinary project teams (Pampel et al., 2013; Cherif et al., 2014; Mahmudlu et al., 2016; Wilkinson et al., 2016).

Furthermore, data collaboration platforms typically offer a unified interface through which users can visualize, modify, and comment on data in an intuitive way. They often integrate with other tools used in the analytical pipeline, such as visualization dashboards, reporting engines, and machine learning platforms. In doing so, they create a holistic environment where both technical and non-technical stakeholders can work with data effectively. However, these advantages come with specific operational and governance challenges (Jacobson et al., 2011; Adepoju et al., 2021). The centralization of data requires a robust infrastructure, compliance with privacy regulations such as GDPR, and clear policies to protect sensitive information. Moreover, organizations must invest in user training and change management processes to ensure that collaborative tools are used correctly, responsibly, and efficiently (Kasse et al., 2018; Tsamados et al., 2022).

In contrast, data federation and virtual database systems adopt a different logic. Rather than bringing data into a centralized platform, these systems leave data in its original location and use a virtual integration layer to make it accessible across various systems. Through this abstraction layer, users can execute queries across multiple sources—be they relational databases, cloud storage systems, APIs, or even unstructured data lakes—without the need to move or duplicate the data (Wu et al., 2010; Gardner et al., 2014). This approach not only reduces the operational burden of data migration but also minimizes the disruption to existing systems and preserves localized data governance models. For actors managing highly distributed or siloed data environments, this represents a powerful way to gain unified insight without sacrificing control (Spiekermann, 2019).

The architecture of these systems supports two main modes of operation: real-time query federation, which retrieves data live from each source system where a query is run, and materialized federation, which creates temporary or scheduled snapshots to improve performance. Many modern solutions combine both approaches in a hybrid architecture, combining the flexibility of real-time access with the speed of cached-data access (Tang et al., 2021). However, these systems come with technical challenges. One of the biggest is schema heterogeneity, when each data source has a different structure or naming convention, making it harder to combine results. Other issues include data synchronization, and network latency, particularly for high-volume or computation-heavy analytical tasks (Bandla, 2024). Security and compliance also remain as important concerns, especially when federated queries span different countries or institutions with their own data protection rules. Ensuring privacy, maintaining auditability, and managing

access rights are essential parts of any well-designed federated system (van Panhuis et al., 2014; Gu et al., 2024).

Ultimately, data collaboration platforms and federated systems represent two ends of a spectrum—centralized collaboration versus decentralized access. The choice between them should not be based on a binary judgment but rather on a nuanced evaluation of organizational needs, technical readiness, regulatory constraints, and strategic objectives. In many cases, the most effective solution will involve a combination of both paradigms, allowing teams to collaborate deeply on selected datasets while also retaining access to broader, distributed data resources. As the data landscape continues to evolve, hybrid models that blend the strengths of collaboration, federation, and virtualization are likely to define the future of enterprise data management. Actors that invest in not only the right tools but also the governance, training, and infrastructure to support them will be best positioned to translate data into insight, action, and competitive advantage (Tian et al., 2016; Petersen et al., 2019; Tang et al., 2021; Andriola, 2023).

This report provides a comprehensive examination of the main methodologies employed in data sharing, analyzing their structural features, functional capacities, and contextual suitability.

Through a comparative analysis of these methodologies, the report aims to elucidate their respective advantages and limitations, as well as to identify the contexts in which each is most effectively deployed. By doing so, it contributes to a deeper understanding of how data sharing strategies can be aligned with organizational goals, regulatory requirements, and technological constraints to foster more intelligent, interoperable, and secure data ecosystems.

2. Methodology and matrix construction

To analyse and compare the different data-sharing approaches, a matrix was developed to examine seven distinct methods, organized across six key analytical categories.

The first category, “WHAT IS”, provides a concise definition of the data-sharing model, providing a brief description of its core mechanism. The second category, “TYPES”, classifies each approach based on the language, model, or framework used. The third category, “HOW IT WORKS”, delves deeper in the explanation of the operational logic behind each method, illustrating how the approach functions in practice. The fourth category, “KEY FEATURES & COMPONENTS”, explores the key components of each approach, aiming to capture their most distinctive relevant characteristics.

Building on this, it was possible to identify the fifth category, “BENEFITS”, in which are identified the specific advantages associated with each approach. Understanding these benefits is essential not only for determining the appropriate domain of application for

each method, but also for assessing where their use may lead to more effective and efficient outcomes. Finally, the last category, “CHALLENGES”, addresses the challenges associated with each system, aiming to highlight not only their weaknesses but also the strengths that may define their potential future applications. Figure 1 summarizes the categories under consideration.

3. Data sharing approaches

3.1 API (Application Programming Interface)

The construction of the API (Application Programming Interface) matrix was approached systematically, aiming to provide a detailed, multi-dimensional framework for understanding how APIs function as a mechanism for data sharing within digital ecosystems. This matrix was developed through the integration of five core analytical dimensions: definition and conceptual framing, types of APIs, operational workflow, structural components, benefits, and challenges. Each dimension captures a critical aspect of API-based data sharing, facilitating a holistic and comparative analysis.

The definition provides a foundational understanding of APIs as standardized sets of protocols and tools designed to facilitate communication between different software systems. Rather than exposing the underlying implementation or source code, APIs act as intermediaries that grant controlled access to data and services. They function as a bridge, fostering interoperability and modular integration across otherwise incompatible platforms.

The typology of APIs was constructed by identifying the most prominent architectural models currently in use: REST (Representational State Transfer), SOAP (Simple Object Access Protocol), and GraphQL. REST, the most adopted framework, uses standard web protocols such as HTTP and supports lightweight data formats like JSON or XML, offering simplicity, scalability, and broad accessibility. In contrast, SOAP employs a stricter XML-based messaging protocol, making it well-suited for enterprise-level applications that require robust security, transaction integrity, and formal service contracts. Finally, GraphQL introduces a more dynamic and client-driven model, allowing users to specify exactly which data they require, thereby optimizing performance and reducing redundancy in data exchange.

The workflow dimension of the matrix outlines the typical operational lifecycle of an API request. This involves a client initiating a request to an API server, which subsequently processes the request, retrieves or modify the necessary data—often through interaction with a backend database—and delivers a structured response back to the client. This workflow underscores the automation and real-time responsiveness that characterize API-based data exchanges.

The structural components of an API are delineated to clarify the architecture that support its functionality. These include well-defined endpoints, which act as access points to specific services; authentication mechanisms, which secure access by validating user credentials or tokens; rate limiting, which restricts the volume of requests a client can issue within a given time to prevent abuse and ensure service stability; and comprehensive documentation, which provides developers with essential guidelines for implementation, integration, and troubleshooting processes.

The benefits of using APIs are multi-faceted. They offer operational efficiency by automating data retrieval and reducing manual interventions. Their inherent flexibility enables disparate applications and systems to interact regardless of their internal architectures. APIs also facilitate real-time access to up-to-date data, which is vital for dynamic environments like financial trading or health monitoring systems. Furthermore, by streamlining data workflows, APIs contribute to reducing the overall costs associated with system maintenance and data management.

Despite the advantages, API-based system present challenges and limitations that are equally critical to acknowledge. Security vulnerabilities may arise from poor authentication practices or insufficient input validation, potentially exposing systems to unauthorized access or data breaches. Scalability and rate-limiting constraints can hinder performance under high-load conditions, frustrating end-users or impairing service delivery. Additionally, version management poses another significant difficulty; updating an API without breaking existing integrations requires careful planning and often involves maintaining backward compatibility. The quality and timeliness of data retrieved via APIs may be compromised if the underlying systems are unsynchronized or suffer from latency issues. Finally, legal and regulatory compliance, particularly in the context of data protection frameworks such as the GDPR, necessitate careful governance to ensure responsible data handling and transparent data ownership.

In sum, the API matrix serves as a comprehensive evaluative framework that examines the operational, architectural, and contextual dimensions of API-based data sharing. By synthesizing theoretical principles with practical considerations, it enables a nuanced understanding of the conditions under which APIs should be adopted, the mechanisms through which they operate, and the strategic purposes they fulfil within digital infrastructures.

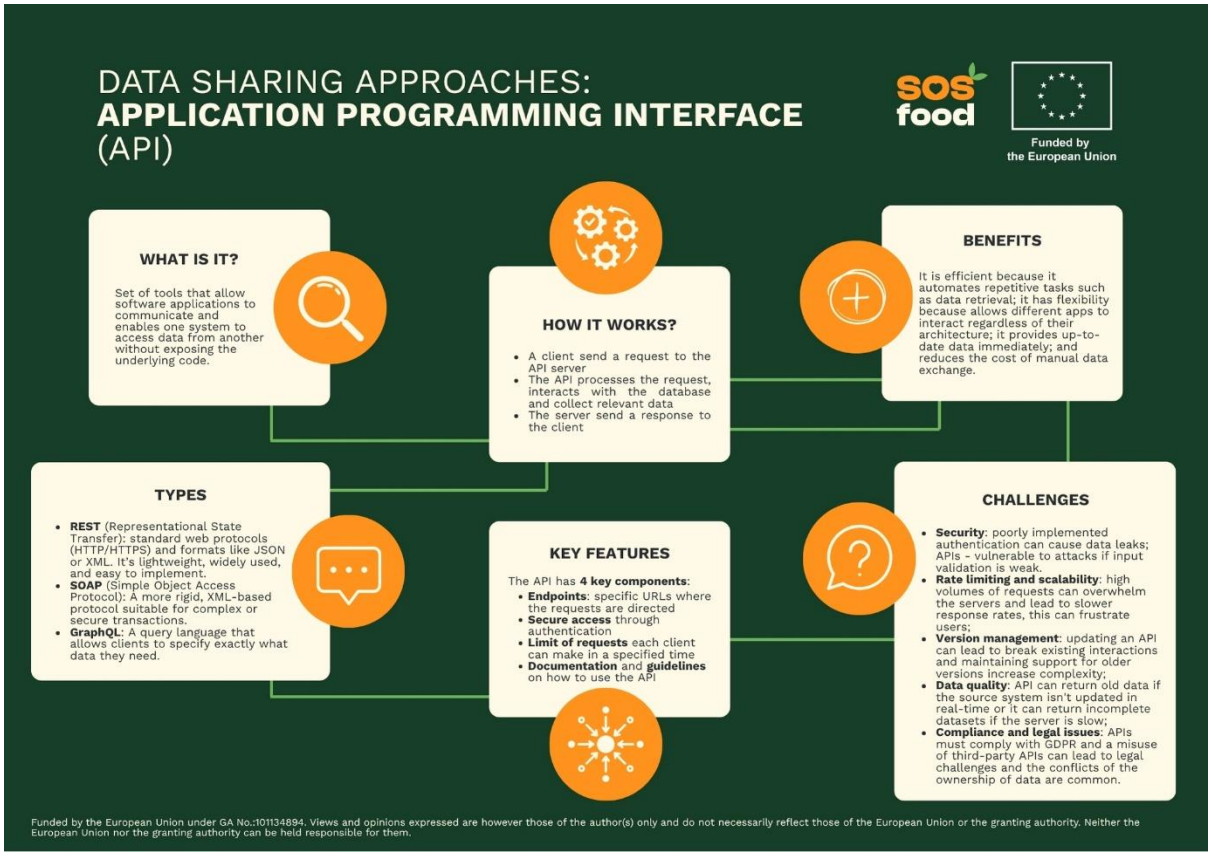

Figure 1. Matrix categories – API

3.2 Data repository

Data repositories play a foundational role in contemporary data management and sharing practices, particularly across academic, governmental, and industrial contexts. These purpose-built infrastructures are designed to support the storage, maintenance, and dissemination of large volumes of digital data. Functioning as centralized platforms, data repositories ensure that information—whether generated through scientific research, institutional operations, or public administration—is preserved in a structured and accessible format. They address both the technical requirements of data storage, such as format compatibility and long-term durability, and the broader goals of accessibility, discoverability, and compliance with data governance standards.

The core function of data repositories is to support both the short- and long-term data sharing by providing reliable and secure storage solutions. Data is typically uploaded in a variety of structured formats and accompanied by detailed metadata, which is critical for contextualizing each dataset. Metadata describes the dataset's origin, structure, purpose, and usage conditions, and often includes details on how the data was collected or generated. This information plays a central role in making data discoverable through search engines or repository-specific queries and allows researchers and other users to interpret, reuse, and cite the data appropriately.

Different types of data repositories serve distinct user groups and objectives. General-purpose repositories are open to all disciplines, making them particularly useful for interdisciplinary research or datasets not aligned with a specific domain repository. In contrast, domain-specific repositories focus on well-defined scientific or professional fields—such as GenBank for genomics or PANGAEA for environmental science—and are optimized for the metadata standards, file formats, and workflows common within those communities. Institutional repositories, typically managed by universities and research centers, host the outputs of affiliated scholars, often as part of broader open access policies or research data management strategies. Governmental repositories, such as Data.gov, are designed to provide public access to datasets produced by public agencies, to support transparency, civic engagement, and evidence-based policymaking. Meanwhile, cloud-based repositories leverage distributed computing resources to offer scalable and flexible storage capabilities, allowing organizations to adapt dynamically to growing data volumes without incurring the costs of managing physical infrastructure.

User interaction with data repositories is carefully structured to ensure both usability and security. Upon uploading their data, users can assign varying levels of access to different stakeholders. Some datasets are made freely available to the public, while others may require restricted access due to privacy concerns, intellectual property rights, or ethical considerations. Repositories employ authentication systems, permission controls, and encryption protocols to protect sensitive information from unauthorized access and to ensure that data handling complies with legal and institutional guidelines.

Beyond basic storage, repositories often support the entire data lifecycle. They provide interfaces for uploading and downloading data, searching metadata, tracking usage statistics, and, in some cases, conducting simple analysis. Many repositories comply with the FAIR principles assuring that data is findable, accessible, interoperable, and reusable, by implementing standardized metadata schemas, persistent identifiers such as DOIs, and compatibility with international data-sharing protocols. These features not only enhance the scientific value of stored data but also promote transparency, reproducibility, and collaboration within and across disciplines.

Despite their significant advantages, data repositories face a range of practical and ethical challenges. Ensuring the security of stored data is crucial, especially in contexts where data may contain personally identifiable information or sensitive commercial insights. Repositories must also navigate complex regulatory landscapes, including regional data protection laws such as the GDPR, which place strict requirements on how data can be stored, accessed, and processed. The quality of metadata remains a persistent issue; datasets that lack sufficient documentation are difficult to find, interpret, or reuse, which undermines the potential for knowledge sharing and innovation.

Financial considerations also present obstacles, as maintaining large-scale data repositories requires ongoing investments in infrastructure, technical support, and

compliance auditing. Furthermore, questions of data ownership and intellectual property are often poorly defined, especially when data is generated collaboratively or funded by multiple stakeholders. This can lead to conflicts over who has the right to access, redistribute, or commercialize the data. Lastly, interoperability remains a significant technical challenge. Different repositories may use different file formats, metadata standards, and access protocols, making it difficult to integrate data across systems without significant manual effort or technical customization.

In conclusion, data repositories are indispensable tools for managing and sharing data in the digital age. They enable centralized, secure, and accessible storage solutions that support data-driven research, policymaking, and innovation. However, to fully realize their potential, ongoing efforts are needed to address the technical, legal, and organizational challenges associated with data security, compliance, metadata quality, cost, ownership, and interoperability.

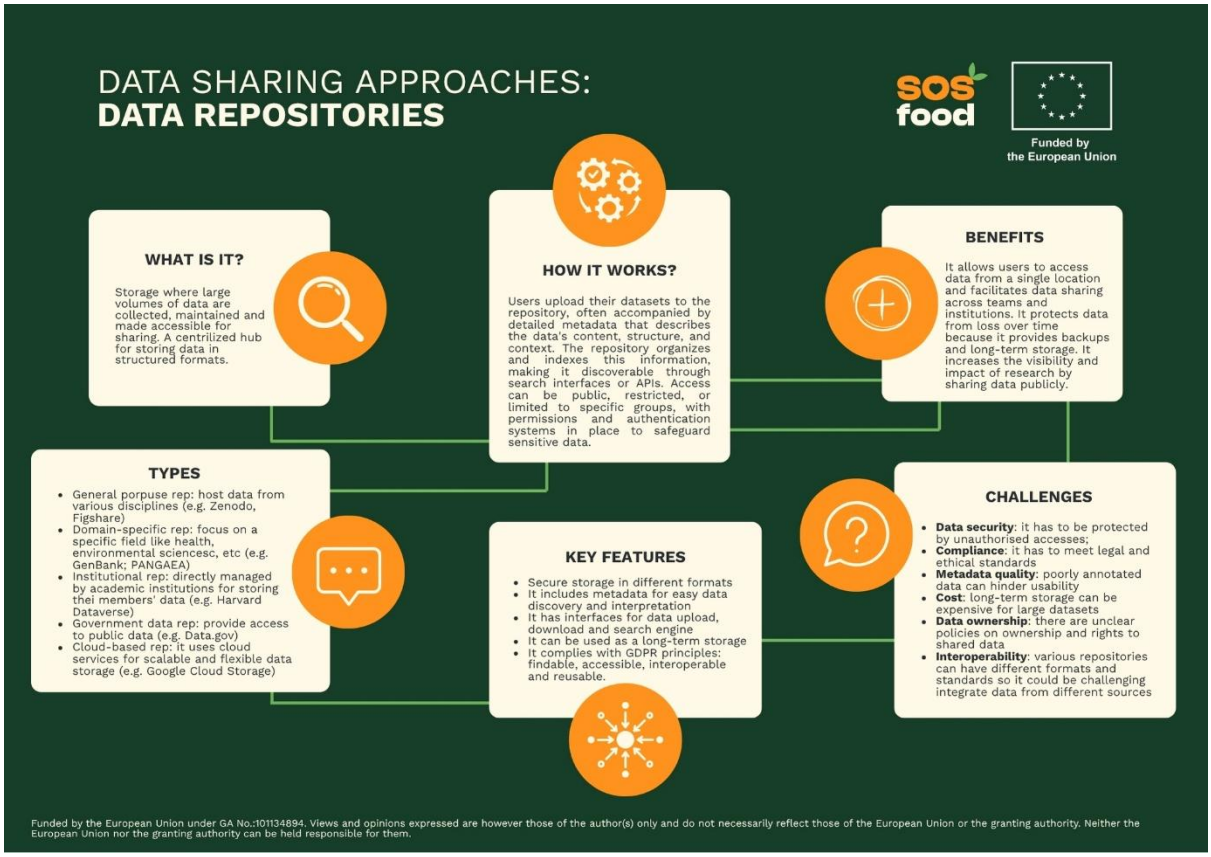

Figure 2. Matrix categories – Data Repositories

3.3 Data Marketplace

Data marketplaces are digital platforms designed to facilitate the exchange of datasets between providers and consumers by enabling the buying, selling, or licensing of data. These platforms serve as commercial hubs where businesses, researchers, and individuals

can access a wide variety of data resources from a centralized location, thereby promoting efficiency, scalability, and accessibility in data acquisition. The rise of data marketplaces reflects the growing recognition of data as a valuable economic asset and aligns with the broader shift toward data-driven decision-making across multiple sectors.

At their core, data marketplaces aggregate datasets from multiple sources and make them available for browsing, purchase, or subscription by end users. These platforms typically include advanced search and filtering tools, as well as detailed metadata and descriptions to help users identify datasets that match their specific needs. Once a dataset is selected, users must review the associated licensing agreements, which define the terms and conditions of use, such as restrictions on redistribution, modification, or commercial exploitation. Transactions are typically completed via secure payment mechanisms, and data may be downloaded or accessed directly through APIs, allowing for seamless integration into analytical workflows and business operations.

Different types of data marketplaces have emerged to address various needs and contexts. Public marketplaces offer open access to data, sometimes free of charge and other times for a fee, supporting transparency and widespread dissemination. In contrast, private marketplaces restrict access to specific users or organizations, typically for the exchange of proprietary or sensitive information. Subscription-based marketplaces allow users to access data through periodic payments, which is especially common for financial or market analytics platforms. Some platforms are sector-specific, serving industries like healthcare, finance, or logistics, and providing datasets aligned with sectoral standards and demands. More recently, decentralized data marketplaces have leveraged blockchain technology to enhance transparency, security, and traceability of data transactions, while reliably recording data provenance and ownership.

One of the main advantages of data marketplaces is their ability to consolidate diverse data sources into a single, navigable environment. This consolidation reduces the time and resources organizations must spend in searching and acquiring relevant data, thereby accelerating decision-making processes and innovation cycles. The availability of customizable datasets allows users to obtain precisely the data needed, enhancing analytical accuracy. For data providers, marketplaces offer a monetization channel for data assets that might otherwise remain underutilized, converting data into a source of revenue and strategic value.

However, operating data marketplaces also introduces challenges and risks. The commercial nature of these platforms raises concerns about data privacy and security, particularly when sensitive or personally identifiable information is involved. The cost of high-quality datasets can be prohibitive, limiting access for smaller organizations or researchers with constrained budgets. Moreover, there is often considerable variability in data quality; not all datasets are curated or validated to the same standard, which can compromise reliability and lead to flawed analyses. Complex licensing arrangements may

impose restrictions that hinder data reuse or redistribution, complicating downstream applications. Interoperability issues can also arise, especially when integrating externally sourced data into legacy systems or heterogeneous digital environments. Finally, users must contend with concerns over trust and transparency, as it is not always easy to assess the credibility of data providers or verify that datasets meet relevant ethical, technical, or legal standards.

In conclusion, data marketplaces represent a powerful tool for the commercialization and dissemination of data, contributing to more efficient and scalable access to information across sectors. While these platforms offer numerous advantages in terms of availability, customization, and system integration, their effective use requires careful attention to legal compliance, data governance, quality assurance, and technical compatibility. As the data economy evolves, the design and regulation of data marketplaces will play a critical role in determining how data is exchanged, valued, and leveraged in the digital era.

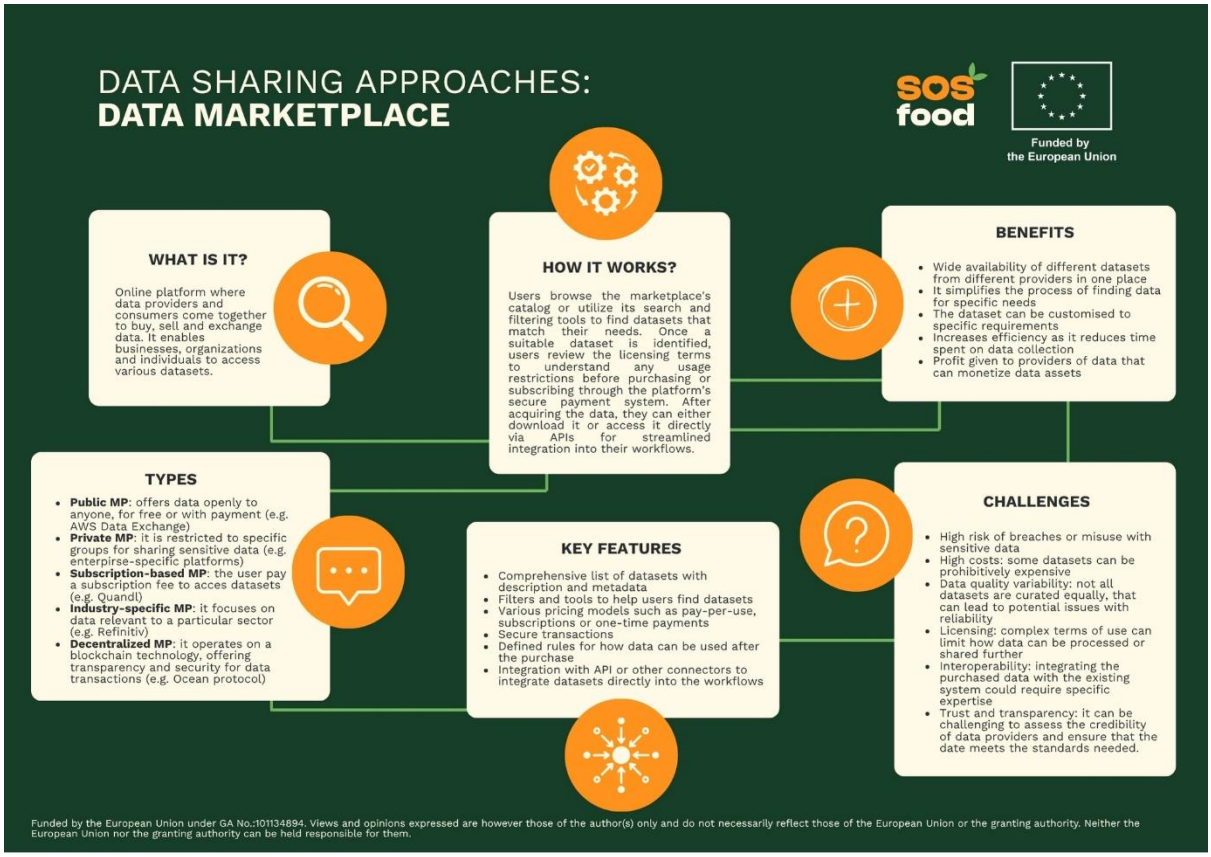

Figure 3. Matrix categories – Data Marketplace

3.4 File sharing services

File sharing services are digital platforms that enable users to upload, store, access, and share files over the internet. These services play a critical role in modern information exchange by facilitating the transfer of various types of digital content—including

documents, images, videos, and software—among individuals, teams, and organizations. Accessible via web interfaces or dedicated applications, they have become essential tools for collaboration, remote work, and digital communication across a wide range of contexts.

At the heart of file sharing services lies the use of centralized or distributed digital storage models. The most ordinary form involves cloud-based systems, where users upload files to remote servers that then generate shareable links or access credentials. Users can define access permissions, such as view-only, editing rights, or download privileges, tailoring availability to the needs of specific collaborators. More advanced platforms offer integrated collaboration tools, such as real-time editing and commenting, enabling multiple users to work on the same document simultaneously and thus enhancing productivity and coordination.

Some services specialize in peer-to-peer (P2P) sharing, where files are exchanged directly between users without relying on a central server, offering decentralization but often sacrificing centralized management and scalability. Enterprise-level platforms further extend these capabilities with features tailored to organizational needs, including role-based access control, integration with internal systems, and audit trails to meet compliance requirements. Additionally, specialized platforms are available to facilitate the transfer of large files that exceed the capacity limits of standard email or messaging systems.

The functionality of file sharing services extends beyond simple file distribution. By enabling remote and multi-device access, these platforms allow users to retrieve and modify documents from virtually any location, supporting the flexibility demanded by contemporary work environments. Security features such as password protection, encryption, and access logs help safeguard shared content. However, the effectiveness of these measures often depends on how diligently users configure privacy settings and access controls.

Due to their convenience and scalability, file sharing services are suitable for a wide range of users, from individuals and small teams to large enterprises managing vast volumes of digital assets. Yet, several challenges accompany their use. A primary concern is the risk of unauthorized access, especially when users fail to configure sharing permissions correctly or when services are targeted by cyberattacks. Many free versions of file sharing platforms impose limits on storage capacity or file size, which can hinder the transfer of large datasets or media files.

Furthermore, ensuring compliance with data protection regulations—such as the GDPR or other jurisdiction-specific frameworks—can be difficult, particularly when files are shared across national borders or with external parties. Additionally, dependency on a stable internet connection can also disrupt access and productivity, particularly in low-connectivity environments. Lastly, as the volume of shared files grows, users and

organizations may face increasing difficulties in managing and organizing content effectively, leading to clutter, duplication, and reduced data traceability.

In conclusion, file sharing services are indispensable tools in the digital era, offering a practical, flexible, and collaborative environment for exchanging information. While they provide significant advantages in terms of accessibility, convenience, and collaborative functionality, their effective use requires careful management of security settings, regulatory compliance, and information organization practices to avoid potential risks and inefficiencies.

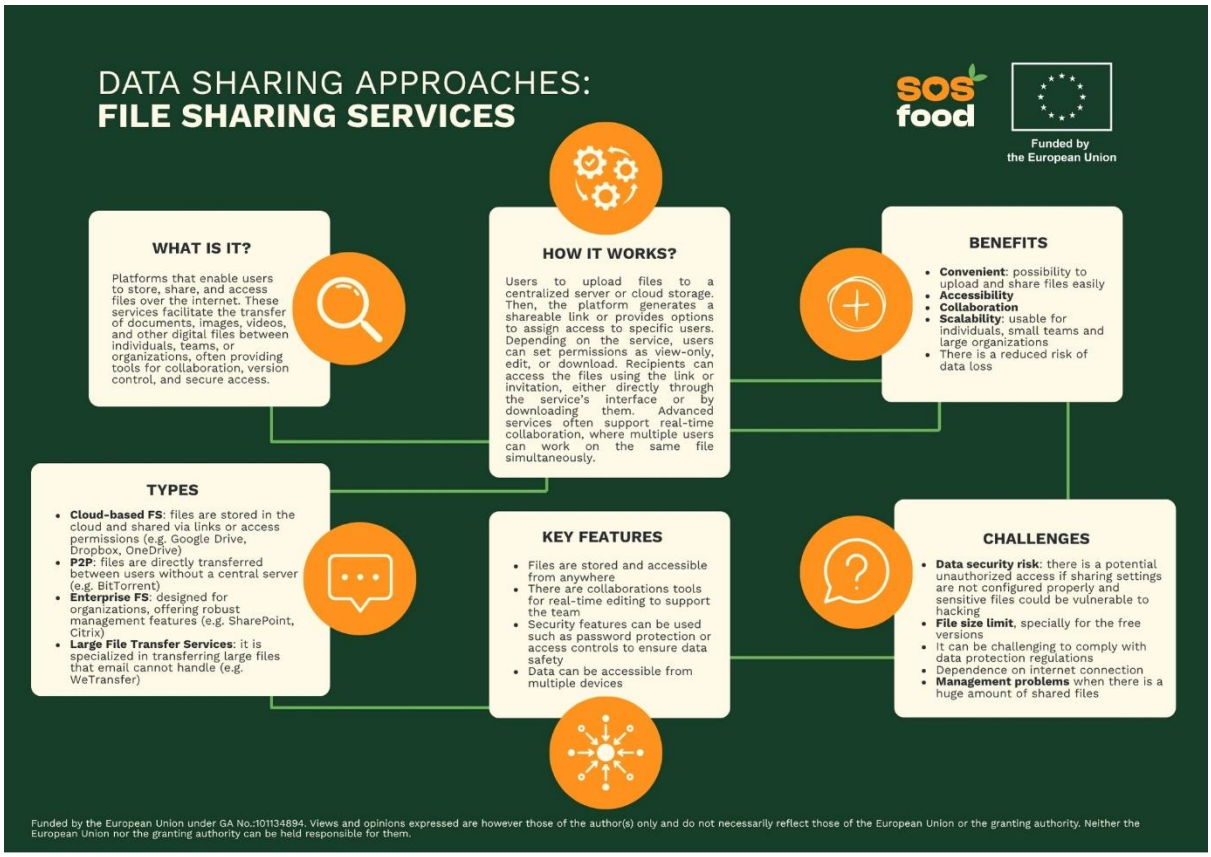

Figure 4. Matrix categories – File Sharing Services

3.5 Peer 2 Peer (P2P) Network

Peer-to-peer (P2P) networks are decentralized communication architectures in which each connected device, or "peer," simultaneously functions as both a client and a server. This configuration eliminates the need for a central server, allowing users to directly exchange data and resources with one another. Each peer in the network can initiate requests, respond to them, and share files, bandwidth, or computational power. The primary feature of P2P systems lies in its distributed nature, which enhances autonomy and reduces reliance on centralized control or infrastructure.

Within a typical P2P network, data is segmented into smaller parts and retrieved from multiple peers concurrently. This multi-source approach not only increases download speed and optimizes bandwidth use but also ensures that no single node is overwhelmed with traffic. After acquiring the file, the recipient peer can then redistribute the data to other users, perpetuating the availability of content within the network.

P2P networks vary in design. Fully decentralized models assign equal roles to all peers, while hybrid models incorporate centralized components for tasks like indexing or search (while preserving decentralized data exchange). Structured P2P systems rely on algorithms to organize peer connections and facilitate efficient data retrieval, whereas unstructured networks lack such organization, which can lead to inefficiencies as the system scales.

P2P networks offer numerous advantages. Their decentralized model reduces infrastructure costs and enhances fault tolerance, since the system gains resources—such as storage and processing power—with each new peer. These characteristics make P2P networks particularly attractive for large-scale file sharing and distributed computing tasks. Additionally, some P2P systems are designed to provide users with a degree of anonymity, supporting private and censorship-resistant communication.

Nevertheless, the decentralized nature of P2P networks also introduces several challenges. Security is a persistent concern, as the lack of centralized oversight can facilitate the spread of malware or unauthorized access to data. Moreover, P2P platforms have often been associated with copyright infringement due to the unregulated sharing of digital content, raising significant legal and ethical issues. The absence of a central authority complicates enforcement and accountability, making it difficult to monitor or manage misuse. Performance variability is another limitation, as the quality of service depends on the availability and behaviour of peers, which may fluctuate over time. In unstructured networks, scalability can lead to inefficiencies in locating data and managing peer connections, especially as the number of participants increases.

In conclusion, P2P networks represent a powerful alternative to traditional centralized systems, offering robust, cost-effective, and scalable solutions for data sharing and resource distribution. However, to fully harness their potential, careful attention must be paid to issues of security, legality, and system performance, particularly as these networks continue to grow in complexity and scale.

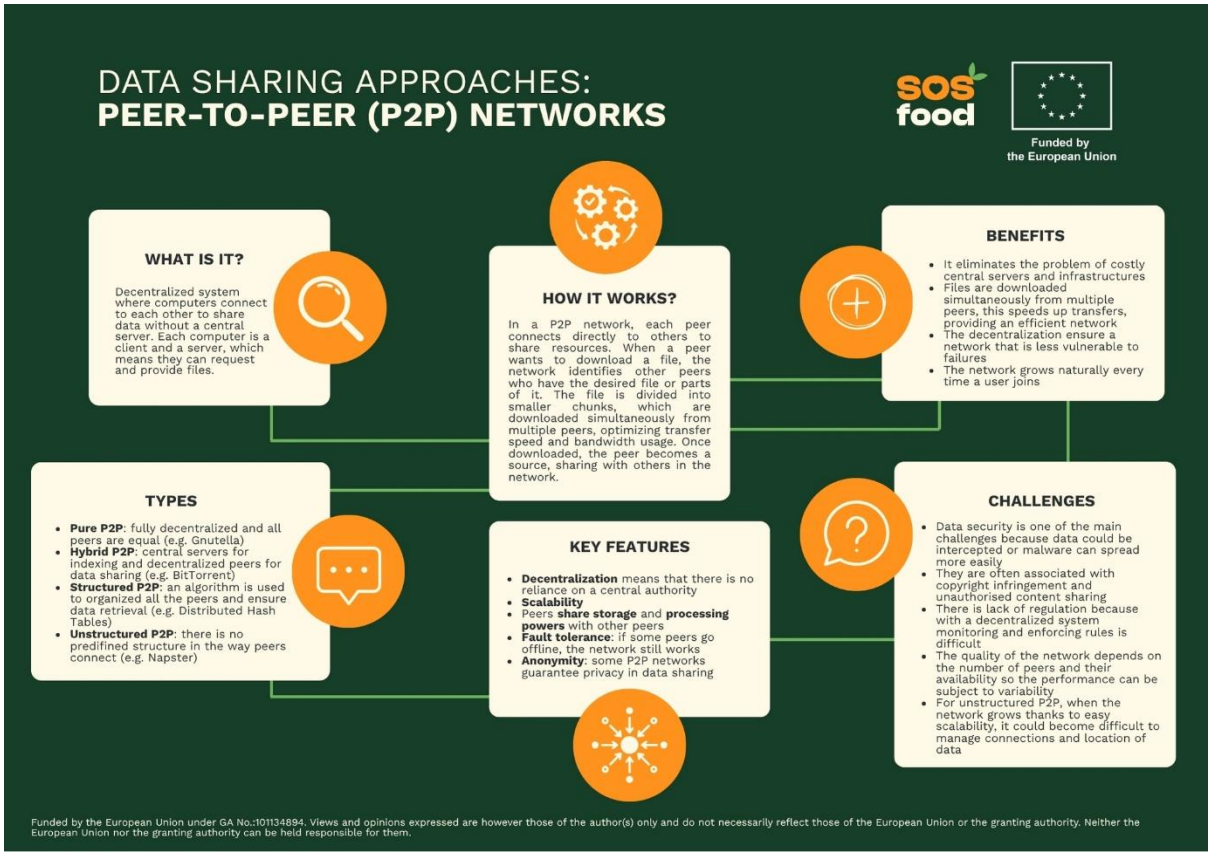

Figure 5. Matrix categories – P2P Networks

3.6 Data Collaboration Platform

Data collaboration platforms are advanced digital environments designed to facilitate the collective work of multiple users on shared datasets in real time. These platforms have become increasingly prevalent in sectors such as scientific research, business intelligence, and software development, where they support data-driven workflows and foster collaborative analysis. By centralizing data access and offering tools for simultaneous interaction, they significantly enhance team productivity, improve the quality of insights generated, and streamline the overall decision-making process.

Typically, data collaboration platforms function by allowing users to upload and organize datasets within a unified interface that supports structured access and governance. Through granular permission settings, contributors can be granted specific capabilities, such as read-only access, editing rights, or administrative control. One of the hallmark features of these platforms is their support for real-time collaboration, whereby multiple users can interact with and modify the same dataset concurrently. Version control and audit trail functionalities are integral to this process, promoting transparency, accountability, and the ability to revert to previous data states when necessary. In addition, robust data governance frameworks are often embedded within these platforms to ensure adherence to privacy laws and institutional policies.

Depending on their deployment model, data collaboration platforms can take various forms. Cloud-based solutions provide flexible, scalable access to computing resources and datasets, accommodating users from different locations without requiring extensive internal infrastructure. On-premises platforms, by contrast, offer organizations tighter control over data security and performance, as they are hosted within the organization's own network environment. Open-source alternatives, often supported by active developer communities, offer customizable and cost-effective options, while industry-specific platforms are tailored to meet the unique data requirements of particular sectors.

Among the key benefits of data collaboration platforms is their capacity to unify data from diverse sources, enabling integrated analysis that draws upon a broader range of inputs. By consolidating access within a centralized system, these platforms reduce data silos, enhance consistency, and foster cross-functional collaboration. Enhanced security features, such as encryption and role-based controls, protect sensitive information, while the inherent scalability of many platforms ensures that growing data volumes and user demands can be accommodated over time.

However, the implementation and use of data collaboration platforms are not without challenges. Organizations must ensure compliance with complex regulatory frameworks such as the GDPR, especially when handling personal or sensitive data. Integrating heterogeneous data sources may also require significant technical expertise and manual effort. Moreover, while some platforms offer powerful functionalities, they may come at a high cost or require specialized training, posing a barrier for smaller teams or institutions. Finally, as collaborative activity scales, performance bottlenecks can arise, potentially affecting the responsiveness of the system and the user experience.

In essence, data collaboration platforms represent a transformative tool in the modern data ecosystem. By enabling secure, real-time, and scalable cooperation around datasets, they empower teams to derive deeper insights, make faster decisions, and operate more efficiently. However, realizing their full potential requires thoughtful implementation, ongoing governance, and attention to the technical and organizational demands they entail.

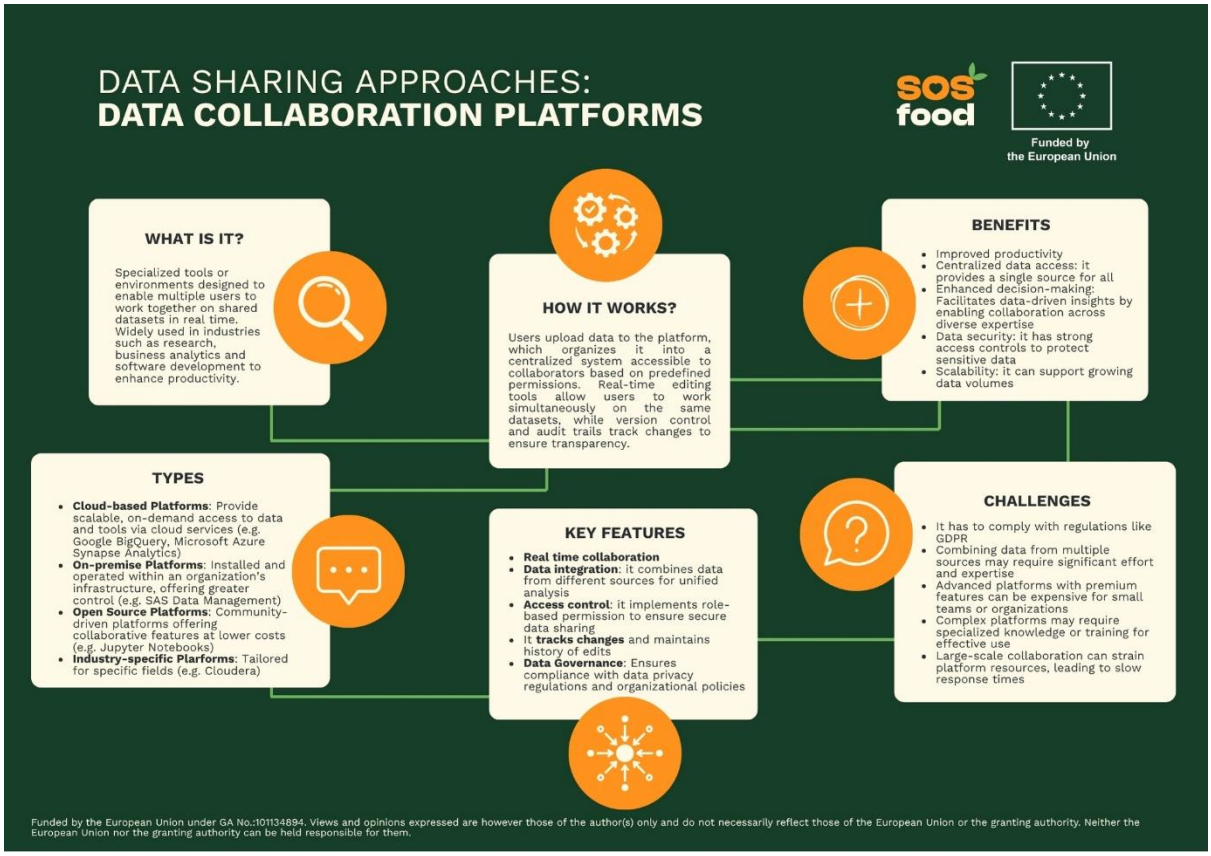

Figure 6. Matrix categories – Data Collaboration Platforms

3.7 Data federations and virtual databases

Data federations and virtual databases represent advanced methodologies for data integration and access that prioritize flexibility and efficiency. These systems allow users to retrieve and analyse data across multiple, distributed sources without necessitating its physical consolidation into a single repository. Built upon the principle of data virtualization, they abstract the underlying complexity of diverse data storage systems to present a unified interface for querying and accessing information. Rather than moving data to a central location, which can be resource-intensive and potentially disruptive, these platforms maintain data in situ—at its original source—thereby preserving existing infrastructure and governance structures while enabling comprehensive analysis across heterogeneous environments.

At the core of these systems is a virtual data layer that serves as an intermediary between users and data sources. This layer is responsible for interpreting user queries, decomposing them into subqueries optimized for each source system, and coordinating the execution of these subqueries either in real time or using cached, materialized snapshots of data, depending on the architecture in use. Query-driven federations perform this process dynamically, retrieving data directly from the source in response to user queries, which is particularly effective for live, interactive analysis. On the other hand,

materialized federations create scheduled or on-demand snapshots of source data to improve performance and reduce latency, making them more suitable for repeated or computationally intensive queries. Hybrid models combine these strategies to optimize both speed and accuracy, allowing for adaptive data processing based on workload demands.

To operate effectively, data federations and virtual databases rely on several foundational technologies and principles. Schema mapping is critical for reconciling differences in data models across systems, facilitating interoperability within a unified query context. These platforms often employ standardized query languages, such as SQL, which allows users to formulate queries across disparate systems using familiar syntax. Real-time data access is another defining feature, as it enables organizations to obtain current insights without waiting for traditional extract, transform, and load (ETL) processes. Furthermore, these platforms are designed with connectivity in mind, supporting integration with a wide array of data sources, including relational databases, cloud-based storage solutions, application programming interfaces (APIs), and even unstructured data environments.

From an organizational perspective, data federations and virtual databases offer considerable benefits. They greatly enhance data accessibility by allowing users to query information across diverse systems without needing to move or duplicate it, which results in substantial time and cost savings. Because they operate atop existing data infrastructures, they minimize the need for disruptive architectural changes and reduce the risks associated with large-scale data migration. Their flexibility allows organizations to adapt quickly to new data sources or changing analytical requirements, all while preserving data governance and lineage.

Nevertheless, challenges persist. Performance can become a critical issue, particularly when executing complex queries over vast and geographically distributed data sources, as latency and bandwidth constraints may impact response times. Inconsistencies in data formats, synchronization lags, or schema mismatches can also lead to analytical discrepancies or incomplete results. Moreover, setting up and maintaining robust schema mappings and metadata alignments often requires a high degree of technical expertise. Ensuring secure and compliant access to all data sources adds another layer of complexity, particularly in contexts where data privacy regulations such as the GDPR must be observed. Finally, certain advanced analytics or machine learning operations may still necessitate centralized data aggregation, as distributed architectures can sometimes limit the scope or performance of such tasks.

In conclusion, data federations and virtual databases provide a powerful framework for integrated, real-time data analysis across decentralized sources. By leveraging virtualization and intelligent query orchestration, they enable organizations to extract insights from diverse datasets without the burdens of traditional data consolidation. However, achieving their full potential requires careful architectural planning, robust

metadata management, and a balanced consideration of performance, security, and regulatory compliance.

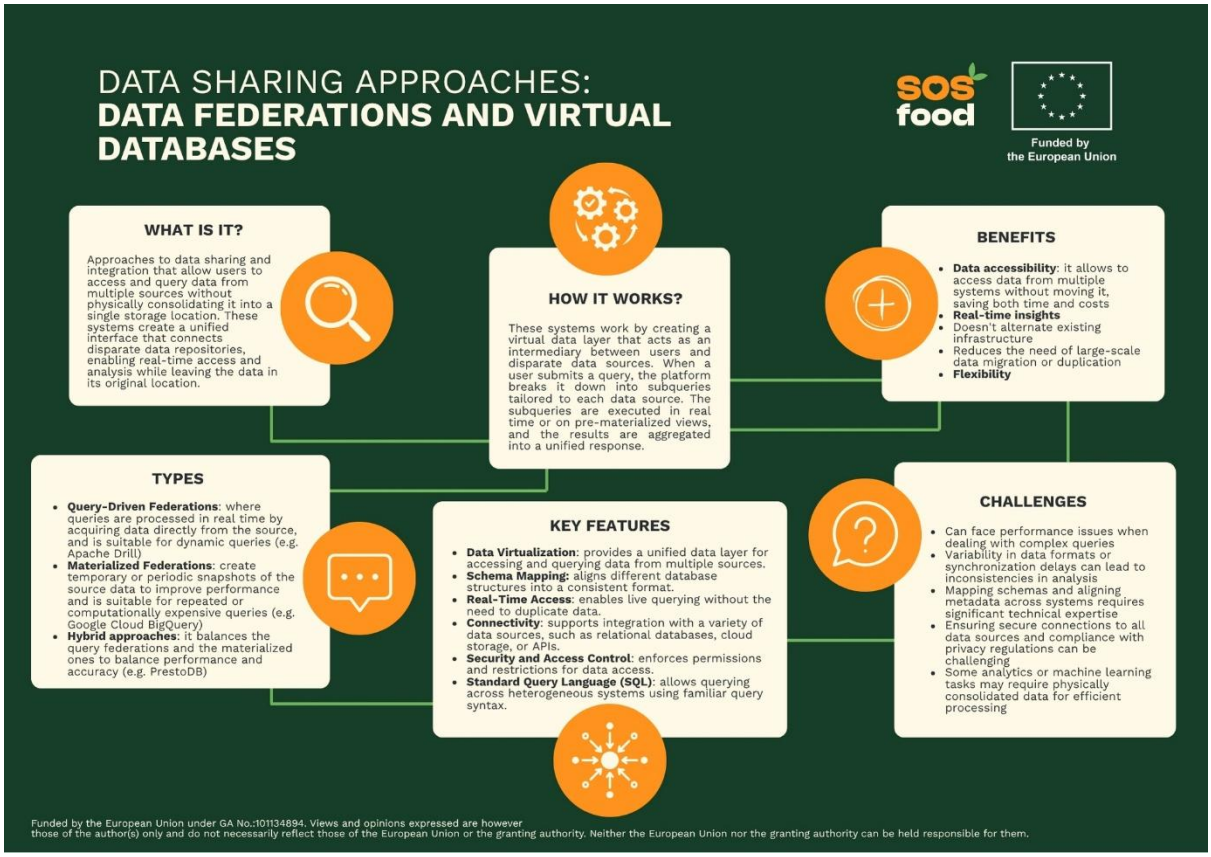

Figure 7. Matrix categories – Data Federations and Virtual Databases

4. Comparison between different approaches

The comparative analysis of APIs and data repositories reveals two fundamentally distinct yet complementary approaches to data sharing within digital ecosystems, each grounded in specific operational logics, technical architectures, and strategic applications. APIs, by design, emphasize real-time, transactional data access through standardized interfaces that facilitate dynamic communication between software systems. Their primary strength lies in enabling interoperability and automation, allowing disparate applications to interact seamlessly and respond instantly to user queries or system events. This makes APIs particularly well-suited to environments where speed, flexibility, and up-to-date information are critical, such as financial trading platforms or health monitoring systems. The modularity of API-based workflows, including features such as endpoint configuration, authentication protocols, and rate limiting, supports scalable and secure integration with external systems. However, this agility comes with challenges, including security vulnerabilities, the complexities of version control, and the risk of service disruptions under high-demand conditions.

In contrast, data repositories operate on a more archival and custodial logic, focusing on the long-term storage, discoverability, and reuse of structured datasets. They prioritize stability, metadata richness, and adherence to governance frameworks, including compliance with FAIR principles and data protection regulations. Unlike APIs, which cater to dynamic exchanges, repositories are optimized for the preservation and contextualization of large, static datasets, often accompanied by detailed metadata to support reproducibility and scholarly citation. This distinction underlines the repositories' strength in facilitating systematic data sharing across research and institutional communities, particularly where transparency, standardization, and longevity are paramount. Nonetheless, repositories face their own set of limitations, including high infrastructure costs, metadata inconsistencies, and ongoing concerns around data ownership and interoperability across platforms.

From this comparison, it is evident that APIs and data repositories address different but overlapping needs: APIs excel in facilitating fast, responsive data flows across heterogeneous systems, while repositories ensure that data is curated, preserved, and accessible for future use. Rather than existing in opposition, these approaches can be viewed as synergistic components of a comprehensive data-sharing strategy, with APIs enabling operational responsiveness and repositories securing the continuity and reliability of digital knowledge. As digital ecosystems grow increasingly complex and data-intensive, the effective coordination between these modalities becomes essential for supporting robust, ethical, and scalable data infrastructures.

Data marketplaces, meanwhile, inject a commercial logic into the ecosystem, positioning data as a tradable commodity and structuring access through licensing, subscriptions, or transactional payments. These platforms enable customizable data acquisition but are burdened by challenges of cost, data quality assurance, and legal complexities surrounding ownership and privacy. File sharing services focus primarily on user-centred accessibility and convenience, enabling ad hoc and collaborative exchanges of digital assets in both personal and organizational contexts. Though efficient for document-level sharing, they often lack the systemic rigor and metadata standards of repositories or the integration depth of APIs.

P2P networks depart most radically from centralized models, instead distributing control and responsibility across participants to create resilient, scalable, and cost-effective environments for data transfer. While P2P systems promote autonomy and may support anonymous or censorship-resistant communication, they are accompanied by heightened risks of security breaches, inconsistent performance, and limited legal accountability. Their decentralized nature complicates the enforcement of access controls, metadata consistency, and collaborative functionality—features that are often essential in institutional or regulated contexts.

Each of these models reflects different trade-offs between control and flexibility, structure and spontaneity, and public utility versus commercial interest. Their effectiveness is largely contingent on contextual factors, including the nature of the data, the intended user base, the required level of security or compliance, and the technical infrastructure available. Rather than being mutually exclusive, these approaches often coexist within complex digital ecosystems, where their distinct features are leveraged in complementary ways to support a wide array of data-driven processes and institutional needs. Understanding their comparative strengths and limitations is therefore essential not only for selecting appropriate tools and platforms but also for shaping sustainable and ethical data governance strategies in an increasingly data-saturated world.

When comparing data collaboration platforms to other approaches such as data marketplaces, file sharing services, P2P networks, and data federations or virtual databases, it becomes evident that each solution embodies a distinct philosophy regarding data accessibility, integration, governance, and user interaction. Data collaboration platforms stand out for their emphasis on structured, real-time, multi-user engagement with shared datasets, offering a cohesive environment that combines centralized access with granular role-based controls. This configuration makes them particularly suited for scenarios requiring transparency, traceability, and active participation from diverse stakeholders, such as scientific research, organizational decision-making, or cross-functional analytics.

In contrast, data marketplaces prioritize the commodification of data, enabling transactional exchanges between producers and consumers but typically lacking the dynamic, co-creative environment that characterizes collaboration platforms. While marketplaces provide extensive access to curated and customizable datasets, they do not inherently support simultaneous user interaction or collaborative editing, focusing instead on data distribution and monetization. Similarly, file sharing services offer mechanisms for sharing documents and files in a relatively unstructured way, often emphasizing convenience and accessibility over integrated analytical workflows. These platforms are effective for ad hoc data exchange but are limited in their ability to support complex, multi-user interactions with evolving datasets, and often lack the governance features intrinsic to data collaboration systems.

P2P networks, on the other hand, are driven by decentralization and autonomy. They excel in scalability and fault tolerance due to their distributed nature, but this very decentralization limits their capacity to enforce consistent access controls, metadata standards, or collaborative functionalities. Their utility is limited in regulated or sensitive environments due to challenges in enforcing access rules or ensuring traceable data usage.

Data federations and virtual databases come closest in function to collaboration platforms by enabling unified access to distributed datasets. However, their primary goal

is query abstraction and integration, not multi-user collaboration. These platforms excel in situations where real-time access to distributed data is essential without physically consolidating it, preserving existing infrastructure and compliance boundaries. Nonetheless, they generally lack the real-time co-editing, workflow management, and collaborative visualization tools found in collaboration platforms. Furthermore, while both approaches can scale and handle complex data environments, federations require sophisticated schema alignment and metadata management, often demanding more technical overhead for setup and maintenance.

In essence, while all these systems contribute to the broader goal of enhancing data accessibility and utility, data collaboration platforms occupy a unique position by merging the technical rigor of centralized governance with the participatory dynamics of real-time, multi-user engagement. Their comparative strength lies in fostering high-trust, high-interaction data environments suited to teams and institutions that require not only access to data but also a structured space in which to work with it collaboratively. As such, they complement but do not substitute the functionalities offered by marketplaces, file sharing platforms, P2P networks, or virtual databases, each of which responds to different technical, organizational, and strategic imperatives.

5. Conclusions

In conclusion, data collaboration platforms and systems such as data federations and virtual databases represent two distinct yet increasingly complementary paradigms for managing, accessing, and leveraging data in contemporary digital ecosystems. Both are responses to the growing demand for scalable, flexible, and secure solutions that enable organizations to generate value from data in an environment characterized by increasing volume, variety, and velocity of information. While they share the overarching objective of improving data accessibility and analytical capability, their approaches, use cases, and architectural foundations differ in meaningful ways, each offering unique advantages and posing specific challenges.

Data collaboration platforms are primarily designed to support real-time, multi-user interaction on shared datasets within a centralized environment. These systems excel in contexts where collaborative insight generation, transparency, and structured governance are essential. By allowing multiple users to access and work on the same datasets simultaneously—often with detailed permission controls and audit trails—they enhance teamwork, reduce redundancy, and support traceable decision-making processes. These platforms are particularly well-suited for organizations that need to consolidate data from different departments or teams and facilitate joint analysis, such as in research consortia, business intelligence units, or cross-functional innovation projects. However, the effectiveness of these systems depends heavily on thoughtful governance frameworks, robust infrastructure, and user training, especially when dealing with sensitive data or complex regulatory environments.

On the other hand, data federations and virtual databases take a fundamentally different approach by decoupling data access from physical data consolidation. Rather than moving or copying data into a centralized repository, these systems allow users to query data directly from its original source through a virtual integration layer. This model is particularly advantageous when data is distributed across multiple departments, geographies, or organizations, and when data sovereignty, compliance, or cost considerations make centralization impractical or undesirable. By abstracting the heterogeneity of source systems and presenting a unified query interface, federated systems significantly reduce data movement, lower integration costs, and preserve the autonomy of underlying data infrastructures. Nevertheless, they also face technical hurdles, including schema mapping complexity, real-time synchronization issues, and potential performance bottlenecks—especially when deployed across large-scale or highly heterogeneous environments.

When evaluated comparatively, it becomes evident that these two paradigms address different aspects of the modern data lifecycle. Data collaboration platforms are optimized for shared editing, real-time decision-making, and version-controlled contribution within controlled settings. Data federations, meanwhile, shine in scenarios where data needs to remain distributed but still be accessed and analysed in a coherent, integrated manner. In many cases, organizations may find that combining elements of both approaches—such as using collaborative platforms for internal teamwork while relying on federated access for external or legacy data sources—offers a more comprehensive and adaptable solution.

Ultimately, the selection or integration of these technologies should be informed by a deep understanding of organizational priorities, technical capacities, regulatory obligations, and strategic goals. The question is not which model is universally superior, but rather which configuration best supports the organization's current and future data needs. As data ecosystems grow more complex, hybrid approaches that blend collaboration, virtualization, and federation will likely become the norm, requiring organizations to invest not only in the right tools, but also in the processes, skills, and governance structures needed to harness their full potential.

This work was supported by the Horizon Europe Framework Programme (HORIZON) [Project number 101134894]. This paper is based on research conducted within the SOSFood (Sustainability Optimization for Secure Food Systems) project. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

DATA SHARING APPROACHES						
APPROACH	WHAT IS IT?	TYPES	HOW IT WORKS	KEY FEATURES & COMPONENTS	BENEFITS	CHALLENGES
API (Application Programming Interface)	Set of tools that allow software applications to communicate and enables one system to access data from another without exposing the underlying code.	1. **REST** (Representational State Transfer): standard web protocols (HTTP/HTTPS) and formats like JSON or XML. It's lightweight, widely used, and easy to implement. 2. **SOAP** (Simple Object Access Protocol): A more rigid, XML-based protocol suitable for complex or secure transactions. 3. **GraphQL**: A query language that allows clients to specify exactly what data they need.	1. A client sends a request to the API server 2. The API processes the request, interacts with the database and collect relevant data 3. The server sends a response to the client	The API has 4 key components: 1. **Endpoints**: specific URLs where the requests are directed 2. **Secure access** through authentication 3. **Limit of requests** each client can make in a specified time 4. **Documentation** and **guidelines** on how to use the API	It is efficient because it automates repetitive tasks such as data retrieval; has flexibility because it allows different apps to interact regardless of their architecture; it provides up-to-date data immediately; reduces the cost of manual data exchange.	1. **Security**: poorly implemented authentication can cause data leaks; APIs – vulnerable to attacks if input validation is weak. 2. **Rate limiting and scalability**: high volumes of requests can overwhelm the servers and lead to slower response rates, this can frustrate users; 3. **Version management**: updating an API can lead to break existing interactions and maintaining support for older versions increase complexity; 4. **Data quality**: API can return old data if the source system isn't updated in real-time or it can return incomplete datasets if the server is slow; 5. **Compliance and legal issues**: APIs must comply with GDPR and a misuse of third-party APIs can lead to legal challenges and the conflicts of the ownership of data are common.
Data Repositories	Storage where large volumes of data are collected, maintained and made accessible for sharing. A centralized hub for storing data in structured formats.	1. **General purpose rep**: host data from various disciplines (e.g. Zenodo, Figshare) 2. **Domain-specific rep**: focus on a specific field like health, environmental sciences, etc (e.g. GenBank; PANGAEA) 3. **Institutional rep**: directly managed by academic institutions for storing their members' data (e.g. Harvard Dataverse) 4. **Government data rep**: provide access to public data (e.g. Data.gov) 5. **Cloud-based rep**: it uses cloud services for scalable and flexible data storage (e.g. Google Cloud Storage)	Users upload their datasets to the repository, often accompanied by detailed metadata that describes the data's content, structure, and context. The repository organizes and indexes this information, making it discoverable through search interfaces or APIs. Access can be public, restricted, or limited to specific groups, with permissions and authentication systems in place to safeguard sensitive data.	1. **Secure storage** in different formats 2. It includes metadata for easy data discovery and interpretation 3. It has interfaces for data upload, download and search engine 4. It can be used as a long-term storage 5. It **complies with GDPR principles**: findable, accessible, interoperable and reusable	It allows users to access data from a single location and facilitates data sharing across teams and institutions. It protects data from loss over time because it provides backups and long-term storage. It increases the visibility and impact of research by sharing data publicly.	1. **Data security**: it has to be protected by unauthorized accesses; 2. **Compliance**: it has to meet legal and ethical standards 3. **Metadata quality**: poorly annotated data can hinder usability 4. **Cost**: long-term storage can be expensive for large datasets 5. **Data ownership**: there are unclear policies on ownership and rights to shared data 6. **Interoperability**: various repositories can have different formats and standards so it could be challenging integrate data from different sources
Data Marketplaces	Online platform where data providers and consumers come together to buy, sell and exchange data. It enables businesses, organizations and individuals to access various datasets.	1. **Public MP**: offers data openly to anyone, for free or with payment (e.g. AWS Data Exchange) 2. **Private MP**: it is restricted to specific groups for sharing sensitive data (e.g. enterprise-specific platforms) 3. **Subscription-based MP**: the user pay a subscription fee to access datasets (e.g. Quandl) 4. **Industry-specific MP**: it focuses on data relevant to a particular sector (e.g. Refinitiv) 5. **Decentralized MP**: it operates on a blockchain technology, offering transparency and security for data transactions (e.g. Ocean protocol)	Users browse the marketplace's catalogue or utilize its search and filtering tools to find datasets that match their needs. Once a suitable dataset is identified, users review the licensing terms to understand any usage restrictions before purchasing or subscribing through the platform's secure payment system. After acquiring the data, they can either download it or access it directly via APIs for streamlined integration into their workflows.	1. Comprehensive list of datasets with description and metadata 2. Filters and tools to help users find datasets 3. Various pricing models such as pay-per-use, subscriptions or one-time payments 4. Secure transactions 5. Defined rules for how data can be used after the purchase 6. Integration with API or other connectors to integrate datasets directly into the workflows	1. Wide availability of **different datasets** from different providers in one place 2. It **simplifies the process** of finding data for specific needs 3. The dataset can be customized to specific requirements 4. **Increases efficiency** as it reduces time spent on data collection 5. Profit given to providers of data that can monetize data assets	1. **High risk of breaches** or misuse with sensitive data 2. **High costs**: some datasets can be prohibitively expensive 3. **Data quality variability**: not all datasets are curated equally, that can lead to potential issues with reliability 4. **Licensing**: complex terms of use can limit how data can be processed or shared further 5. **Interoperability**: integrating the purchased data with the existing system could require specific expertise 6. **Trust and transparency**: it can be challenging to assess the credibility of data providers and ensure that the date meets the standards needed.

File Sharing Services	Platforms that enable users to store, share, and access files over the internet. These services facilitate the transfer of documents, images, videos, and other digital files between individuals, teams, or organizations, often providing tools for collaboration, version control, and secure access.	1. **Cloud-based FS**: files are stored in the cloud and shared via links or access permissions (e.g. Google Drive, Dropbox, OneDrive) 2. **P2P**: files are directly transferred between users without a central server (e.g. BitTorrent) 3. **Enterprise FS**: designed for organizations, offering robust management features (e.g. SharePoint, Citrix) 4. **Large File Transfer Services**: it is specialized in transferring large files that email cannot handle (e.g. WeTransfer)	Users to upload files to a centralized server or cloud storage. Then, the platform generates a shareable link or provides options to assign access to specific users. Depending on the service, users can set permissions as view-only, edit, or download. Recipients can access the files using the link or invitation, either directly through the service's interface or by downloading them. Advanced services often support real-time collaboration, where multiple users can work on the same file simultaneously.	1. Files are stored and accessible from anywhere 2. There are collaborations tools for real-time editing to support the team 3. Security features can be used such as password protection or access controls to ensure data safety 4. Data can be accessible from multiple devices	1. **Convenient:** possibility to upload and share files easily 2. **Accessibility** 3. **Collaboration** 3. **Scalability:** usable for individuals, small teams and large organizations 4. There is a reduced risk of data loss	1. **Data security risk**: there is a potential unauthorized access if sharing settings are not configured properly and sensitive files could be vulnerable to hacking 2. **File size limit**, especially for the free versions 3. It can be challenging to comply with data protection regulations 4. Dependence on internet connection 5. Management problems when there is a huge amount of shared files
Peer-to-Peer (P2P) Networks	Decentralized system where computers connect to each other to share data without a central server. Each computer is a client and a server, which means they can request and provide files.	1. **Pure P2P**: fully decentralized and all peers are equal (e.g. Gnutella) 2. **Hybrid P2P**: central servers for indexing and decentralized peers for data sharing (e.g. BitTorrent) 3. **Structured P2P**: an algorithm is used to organized all the peers and ensure data retrieval (e.g. Distributed Hash Tables) 4. **Unstructured P2P**: there is no predefined structure in the way peers connect (e.g. Napster)	In a P2P network, each peer connects directly to others to share resources. When a peer wants to download a file, the network identifies other peers who have the desired file or parts of it. The file is divided into smaller chunks, which are downloaded simultaneously from multiple peers, optimizing transfer speed and bandwidth usage. Once downloaded, the peer becomes a source, sharing with others in the network.	1. **Decentralization** means that there is no reliance on a central authority 2. **Scalability** 3. Peers share storage and processing powers with other peers 4. **Fault tolerance**: if some peers go offline, the network still works 5. **Anonymity**: some P2P networks guarantee privacy in data sharing	1. It eliminates the problem of costly central servers and infrastructures 2. Files are downloaded simultaneously from multiple peers, this speeds up transfers, providing an efficient network 3. The decentralization ensure a network that is less vulnerable to failures 4. The network grows naturally every time a user joins	1. **Data security** is one of the main challenges because data could be intercepted or malwares can spread more easily 2. They are often associated with **copyright infringement** and **unauthorized content sharing** 3. There is **lack of regulation** because with a decentralized system monitoring and enforcing rules is difficult 4. The **quality** of the network depends on the number of peers and their availability so the performance can be subject to variability 5. For unstructured P2P, when the network grows thanks to easy scalability, it could become **difficult to manage connections** and **location** of data
Data Collaboration Platforms	Specialized tools or environments designed to enable multiple users to work together on shared datasets in real time. Widely used in industries such as research, business analytics and software development to enhance productivity.	1. **Cloud-based Platforms**: Provide scalable, on-demand access to data and tools via cloud services (e.g. Google BigQuery, Microsoft Azure Synapse Analytics) 2. **On-premise Platforms**: Installed and operated within an organization's infrastructure, offering greater control (e.g. SAS Data Management) 3. **Open Source Platforms**: Community-driven platforms offering collaborative features at lower costs (e.g. Jupyter Notebooks) 4. **Industry-specific Platforms**: Tailored for specific fields (e.g. Cloudera)	Users upload data to the platform, which organizes it into a centralized system accessible to collaborators based on predefined permissions. Real-time editing tools allow users to work simultaneously on the same datasets, while version control and audit trails track changes to ensure transparency.	1. **Real time collaboration** 2. **Data integration**: it combines data from different sources for unified analysis 3. **Access control**: it implements role-based permission to ensure secure data sharing 4. It **tracks changes** and maintains history of edits 5. **Data Governance**: Ensures compliance with data privacy regulations and organizational policies	1. **Improved productivity** 2. **Centralized data access**: it provides a single source for all 3. **Enhanced decision-making**: Facilitates data-driven insights by enabling collaboration across diverse expertise 4. **Data security**: it has strong access controls to protect sensitive data 5. **Scalability**: it can support growing data volumes	1. It has to comply with regulations like **GDPR** 2. Combining data from multiple sources may require **significant effort** and expertise 3. Advanced platforms with premium features can be **expensive** for small teams or organizations 4. Complex platforms may require **specialized knowledge** or training for effective use 5. Large-scale collaboration can strain platform resources, leading to **slow response times**

Data Federations and Virtual Databases	Approaches to data sharing and integration that allow users to access and query data from multiple sources without physically consolidating it into a single storage location. These systems create a unified interface that connects disparate data repositories, enabling real-time access and analysis while leaving the data in its original location.	1. **Query-Driven Federations**: where queries are processed in real time by acquiring data directly from the source, and is suitable for dynamic queries (e.g. Apache Drill) 2. **Materialized Federations**: create temporary or periodic snapshots of the source data to improve performance and is suitable for repeated or computationally expensive queries (e.g. Google Cloud BigQuery) 3. **Hybrid approaches**: it balances the query federations and the materialized ones to balance performance and accuracy (e.g. PrestoDB)	These systems work by creating a virtual data layer that acts as an intermediary between users and disparate data sources. When a user submits a query, the platform breaks it down into subqueries tailored to each data source. The subqueries are executed in real time or on pre-materialized views, and the results are aggregated into a unified response.	1. **Data Virtualization**: provides a unified data layer for accessing and querying data from multiple sources. 2. **Schema Mapping**: aligns different database structures into a consistent format. 3. **Real-Time Access**: enables live querying without the need to duplicate data. 4. **Connectivity**: supports integration with a variety of data sources, such as relational databases, cloud storage, or APIs. 5. **Security and Access Control**: enforces permissions and restrictions for data access. 6. **Standard Query Language** (SQL): allows querying across heterogeneous systems using familiar query syntax.	1. **Data accessibility**: it allows to access data from multiple systems without moving it, saving both time and costs 2. **Real-time insights** 3. Doesn't alternate existing infrastructure 4. Reduces the need of large-scale data migration or duplication 5. **Flexibility**	1. Can face performance issues when dealing with complex queries 2. Variability in data formats or synchronization delays can lead to **inconsistencies** in analysis 3. Mapping schemas and aligning metadata across systems requires **significant technical expertise** 4. Ensuring secure connections to all data sources and compliance with privacy regulations can be challenging 5. Some analytics or machine learning tasks may require physically consolidated data for efficient processing

References

1. Adepoju, A. H., Austin-Gabriel, B., Hamza, O., & Collins, A. (2021). A conceptual model for centralized data platforms to enhance decision-making and optimize cross-functional collaboration. Open Access Research Journal of Science and Technology, 2(1), 23—40. https://doi.org/10.53022/oarjst.2021.2.1.0049

2. Andriola, T. (2023). Data as Strategy: How Platforms Solve Problems and Ecosystems can Transform the Industry. Telehealth and Medicine Today, 8(2). https://doi.org/10.30953/thmt.v8.395

3. Bandla, C. (2024). Distributed Database Architectures for Federated Medical Training. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 4(2), 633—641. https://doi.org/10.48175/IJARSCT-22774

4. Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world. The MIT Press. https://doi.org/10.7551/mitpress/9963.001.0001

5. Cherif, A., Imine, A., & Rusinowitch, M. (2014). Practical access control management for distributed collaborative editors. Pervasive and Mobile Computing, 15, 62—86. https://doi.org/10.1016/j.pmcj.2013.09.004 ù

6. Gu, Z., Corcoglioniti, F., Lanti, D., Mosca, A., Xiao, G., Xiong, J., & Calvanese, D. (2024). A Systematic Overview of Data Federation Systems. Semantic Web 15 (2024) 107—165. https://doi.org/10.3233/SW-223201

7. Halevy, A. Y., Rajaraman, A., & Ordille, J. (2006). Data integration: The teenage years. Communications of the ACM, 51(9), 9—13. https://doi.org/10.1145/1246748.1246752

8. Jacobson, D., Brail, G., & Woods, D. (2011). APIs: A strategy guide. O'Reilly Media.

9. Kasse, J. P., Xu, L., deVrieze, P., & Bai, Y. (2018). The need for compliance verification in collaborative business processes. In L. M. Camarinha-Matos, H. Afsarmanesh, & D. Antonelli (Eds.), Collaboration in a hyperconnected world (pp. 217—229). Springer. https://doi.org/10.1007/978-3-319-99127-6_19

10. Kitchin, R. (2014). The data revolution: big data, open data, data infrastructures and their consequences. SAGE Publications Ltd. https://doi.org/10.4135/9781473909472

11. Kowalczyk, S., & Shankar, K. (2011). Data sharing in the sciences. Annual Review of Information Science and Technology, 45(1), 247—294. https://doi.org/10.1002/aris.2011.1440450113

12. Leonelli, S. (2016). Data-centric biology: A philosophical study. University of Chicago Press.

13. Loshin, D. (2011). The practitioner's guide to data quality improvement. Morgan Kaufmann OMG Press.

14. Mahmudlu, R., Hartog, J.D., & Zannone, N. (2016). Data Governance & Transparency for Collaborative Systems.

15. Mbuagbaw, L., Foster, G., Cheng, J., & Thabane, L. (2017). Challenges to complete and useful data sharing. Trials, 18(1), 71. https://doi.org/10.1186/s13063-017-1816-8

16. Michener, W. K. (2015). Ecological data sharing. Ecological Informatics, 29, 33—44. https://doi.org/10.1016/j.ecoinf.2015.06.010

17. Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., & Dierolf, U. (2013). Making Research Data Repositories Visible: The re3data.org Registry. PLOS ONE, 8(11), e78080. https://doi.org/10.1371/journal.pone.0078080

18. Parashar, M., Simonet, A., Rodero, I., Ghahramani, F., Agnew, G., & Jantz, R. (2020). The Virtual Data Collaboratory: A Regional Cyberinfrastructure for Collaborative Data-Driven Research. Computing in Science & Engineering, 22(3), 79—92. https://doi.org/10.1109/MCSE.2019.2908850

19. Petersen, S. A., Pourzolfaghar, Z., Alloush, I., Ahlers, D., Krogstie, J., & Helfert, M. (2019). Value-Added Services, Virtual Enterprises and Data Spaces Inspired Enterprise Architecture for Smart Cities. In L. M. Camarinha-Matos, H. Afsarmanesh, & D. Antonelli (Eds.), Collaborative Networks and Digital Transformation (pp. 393—402). Springer. https://doi.org/10.1007/978-3-030-28464-0_34

20. Spiekermann, M. (2019). Data Marketplaces: Trends and Monetisation of Data Goods. Intereconomics 54, 208—216. https://doi.org/10.1007/s10272-019-0826-z

21. Tang, C., Wang, B., Wu, H., Wang, Z., Li, Y., Channapattan, V., Luo, Z., Kabra, R., Ghosh, M., Navadiya, N.K., & Mishra, P. (2021). Hybrid-Cloud SQL Federation System at Twitter. European Conference on Software Architecture.

22. Tian, Y., Özcan, F., Zou, T., Goncalves, R., & Pirahesh, H. (2016). Building a hybrid warehouse: Efficient joins between data stored in HDFS and enterprise warehouse. ACM Transactions on Database Systems, 41(4), Article 21. https://doi.org/10.1145/2972950

23. Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: Key problems and solutions. AI & Society, 37. https://doi.org/10.1007/s00146-021-01154-8

24. van Panhuis, W.G., Paul, P., Emerson, C. et al. (2014). A systematic review of barriers to data sharing in public health. BMC Public Health 14, 1144. https://doi.org/10.1186/1471-2458-14-1144

25. Wickramasinghe, N., & Bodendorf, F. (Eds.). (2020). Delivering Superior Health and Wellness Management with IoT and Analytics. Springer International Publishing. https://doi.org/10.1007/978-3-030-17347-0

26. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

27. Wu, S., Xiao, T., Fan, W., & Cui, B. (2010). Collaborative development system based on federation integration. In Proceedings of the 2010 First International Conference on Networking and Distributed Computing (pp. 368—372). IEEE. https://doi.org/10.1109/ICNDC.2010.77

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA).

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

All partners declare to comply with the SOSFOOD non-discrimination commitment.