Federated Learning for Data Spaces: a Privacy-Enhancing Strategy based on Data Visiting

Manlio Bacco[1], Margherita Di Leo[2], Albana Kona[1], Mattia Santoro[3], Paolo Mazzetti[3]

[1]European Commission, Joint Research Centre (JRC), Ispra (VA), Italy
[2]Arcadia SIT, under contract with the European Commission, JRC
[3]Institute of Atmospheric Pollution Research—National Research Council of Italy (CNR-IIA), 50019 Sesto Fiorentino, Italy

***Abstract*—This work explores the paradigm of data visiting that, through privacy-enhancing technologies, shows the potential to access and use data otherwise inaccessible. Building on the ongoing EU initiative to design, implement, and run sectorial data spaces, we consider federated learning as one the most promising approaches for the objective above. We propose a domain-agnostic strategy that can be extended and adapted to different needs. We conclude by analysing the limitations and challenges of the approach we propose.**

***Index Terms*—data visiting, privacy-enhancing technologies, federated learning, data spaces**

I. Introduction

In 2020, the European Commission (EC) unveiled the European Strategy for Data, a vision advocating for the establishment of specialized data ecosystems, namely *data spaces*, across various sectors. These ecosystems are designed to facilitate the sharing and reutilization of diverse data types—ranging from governmental and private sector data to research outputs and information contributed by citizens- with the ultimate objective of establishing a single market for data. To guarantee the successful establishment of data spaces, the strategy is built upon a robust foundation of overarching legislative actions, such as the Data Governance Act [1], the Data Act [2] and funding programs such as Digital Europe [3]. Collectively, these efforts will play a pivotal role in facilitating the rollout of common European data spaces, ultimately contributing to the dynamism and competitiveness of the European data economy [4].

Central to this strategy is the adoption of the 'data visiting' paradigm, also known as *compute to data*, which refers to the possibility of using data where they are, without moving them to a different location before use. Hence, it is the algorithm (or, more generally, the computational process) that moves where the data are stored, and not vice versa (i.e., *data to compute*) as common nowadays. This approach is particularly useful when dealing with large volumes of data that are impractical or inefficient to transfer due to their size, privacy concerns, or regulatory restrictions [5]. In traditional data analysis, data is transferred over networks to the analyst's local environment, which can be time-consuming and pose security risks. With data visiting, operations are performed on the data where it resides. Data visiting supports FAIR (Findable, Accessible, Interoperable, Reusable) data sharing principles [6] by allowing users to access and work with data from multiple sources without duplicating or moving it, thus saving time, preserving bandwidth, and enhancing security compared to traditional data analysis.

Data visiting is gaining popularity in fields like healthcare and scientific research where data privacy and scale are critical concerns. It is also increasingly common practice in federated systems. A federated data system involves establishing governance authorities within each data domain, such as customer data/production data, and business units, such as marketing/customer services. Data spaces implement by design the federated data governance paradigm. They are distributed systems defined by a governance framework that facilitates secure and trustworthy data transactions between participants, emphasising trust and data sovereignty [7].

The federated nature of data spaces aligns well with the use of Privacy-Enhancing Technologies (PETs) [8], which can be applied to protect sensitive (such as personal, business, and so on) data while still enabling collaborative data use and analysis. In data spaces, data intermediaries are organisations with a variety of roles [9], [10], such as intermediation of technical solutions and infrastructures, making use of legal constructs (e.g., data trusts), or other collective governance mechanisms (e.g., data cooperatives). In this work, technical solutions and infrastructures are of interest because they offer building blocks to use data in a privacy-preserving fashion.

This work is organised as follows. In Section II, we survey the state of the art when it comes to PETs and focus on exemplary use cases in the literature. In Section III, we provide a short analysis of the pros and cons of centralised and distributed Machine Learning (ML) approaches, considering privacy as key requirement. Then, we focus our attention on Federated Learning (FL) in Section IV. FL can be a key strategy for data visiting, and we present how it works and discuss its advantages and disadvantages. In Section V, we propose a domain-agnostic architecture for the use of data visiting strategies, especially in data spaces, and we present its potential use in a forestry use case. Finally, in Section VI we look at challenges and limitations of the approach we propose, and then we conclude in Section VII.

corresponding author: M. Bacco manlio.bacco@ec.europa.eu

II. State of the art

In this section, we focus on PETs and exemplary use cases in the literature. As anticipated in Section I, the federated governance model of data spaces provides an ideal framework for implementing PETs. In data spaces, each domain can tailor PETs to specific needs, data, and use case requirements. Key technologies that can be integrated into data space architectures to enable privacy-preserving data use, collaboration, and analysis include the following ones, among others [8]:

- Secure Multi-Party Computation: it allows multiple parties to jointly compute a function over their inputs without revealing the inputs to each other. This enables collaborative data analysis without exposing the underlying sensitive data.
- Differential Privacy: it can be used to add noise to data outputs, allowing data to be shared or analysed while providing strong privacy guarantees for the individuals represented in the data.
- Zero-Knowledge Proofs: they allow one party to prove that a statement is true, without revealing any additional information. This can be used to verify data or model properties without exposing the underlying data.
- Federated Learning (FL): a distributed solution fully described in Section IV, it implements the compute-to-data paradigm.

PETs, for instance FL, are highly relevant and beneficial in e.g., healthcare use cases [11]. FL has been successfully used for rare disease detection [12], leveraging data from multiple sources to train accurate AI models while preserving patient privacy. FL proves to be well-suited in such a scenario for several reasons. Firstly, rare diseases have small sample sizes, making it difficult to train accurate AI models. FL aggregates data from multiple institutions without moving them, allowing models to be trained on larger and heterogeneous datasets [12]. Secondly, patient data are not moved nor shared by healthcare institutions because of privacy concerns. FL addresses this by only sharing model updates (not data), still allowing collaborative model training [13], [14]. Thirdly, FL has shown good scalability, meaning that collaboration can be carried out also in large settings, as in [13], showing how data of rare cancer boundary detection from 29 institutions has ensured meaningful results for rare diseases.Anyway, uneven data distribution, common across different institutions, can be a challenge for FL, but novel techniques improve the achievable accuracy [14].

Another interesting use case for FL is forestry [15], [16]. Take the case of the EU, in which forest field data are collected by National Forest Inventories (NFIs) and used for a wide range of purposes, e.g. to assess the health and condition of forests, to monitor changes in forest cover, and to plan sustainable forest management practices; to obtain information on the diversity of tree species, ecological processes and biodiversity conservation; and so on. Governments and policy-makers use NFI data to develop policies related to forest conservation and natural resources management, while scientific research uses it to understand the dynamics of forests and their ecosystems. Looking at data collection, we emphasise how plot data are manually collected in the field, making the process costly and poorly feasible in vast areas; anyway, high-quality data like manually collected ones are crucial for model calibration, thus they hold great value. NFI data are considered sensitive for several reasons, thus unlikely to be publicly available. To add to the complexity, terms and conditions of use vary according to the country. Deriving information at a regional level is a critical objective, but it requires the integration of field measurements with remote sensing data. For these reasons, it is important to access manual measurements to calibrate models based on remote imagery. At present, common solutions see data being degraded before being shared, coordinates of the plots being shifted, and noise being added. However, by doing so, the data utility for modellers drops dramatically. This is to say that there is an obvious trade-off between privacy and utility [17]: the more privacy is preserved, the less utility (accuracy) is achieved in modelling. Such a trade-off stands also for FL, but models can be trained on unobfuscated NFI data because only model updates (not raw data) are shared among participants. As in the healthcare case, the aggregation of data from heterogeneous sources has a beneficial effect on the quality of the models, thus overcoming the limitations of small plot-level datasets collected in the field. Such collaborative frameworks also foster a collaborative model development across countries and institutions, leveraging the diverse forestry-related expertise, thus leading to more robust and accurate models for assessing forest health, carbon stocks, and other key metrics. FL also has the potential to accommodate the different terms and conditions of data access and usage across countries, indeed enabling a harmonized modelling framework despite the heterogeneous data landscape.

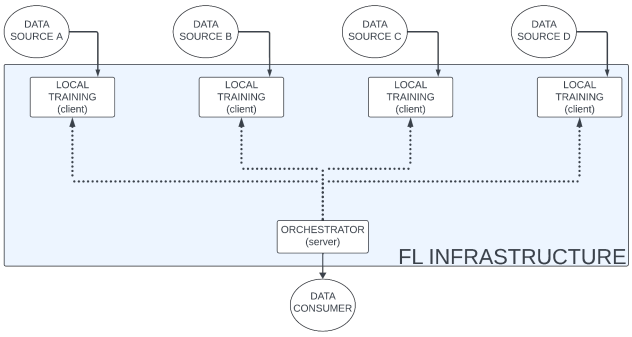

Fig. 1: Federated Learning: the data consumer deploys an orchestrator that controls a set of clients performing local training on data from several providers. Once completed, the results of the collaborative training are returned to the consumer in the form of a trained learning model.

We conclude the analysis of the state of the art by considering the case described in [18], in which the authors propose a software solution for the deployment of FL in data spaces. Using parking data as a use case, they instantiate a data space

TABLE I: List (not exhaustive) of available FL frameworks.

Name	Code repository
CrypTen	https://github.com/facebookresearch/CrypTen
FATE	https://github.com/FederatedAI/FATE
FedN	https://github.com/scaleoutsystems/fedn
Flower	https://github.com/adap/flower
FLSim	https://github.com/facebookresearch/FLSim
LEAF	https://github.com/TalwalkarLab/leaf
IBM FL	https://github.com/IBM/federated-learning-lib
NVIDIA FLARE	https://github.com/NVIDIA/NVFlare
OpenFL	https://github.com/securefederatedai/openfl
PySyft	https://github.com/OpenMined/PySyft
Substra	https://github.com/substra
TFF	https://github.com/google-parfait/tensorflow-federated
FedML	https://github.com/FedML-AI

infrastructure and the components they developed to run FL. Reference [18] shares with this work the aim of a better understanding of the potential uses of FL, technical feasibility, and practical implications of FL in data spaces, concluding that especially economic feasibility remains an open issue. In the following section, we discuss centralised solutions as opposed to distributed ones.

III. Centralised, Distributed On-Site, and Federated Solutions

Before moving to a deeper analysis of FL, we discuss the main differences among centralised, distributed on-site, and federated solutions for ML. We focus on privacy as key requirement guiding the discussion, and we refer the reader to Fig. 2 to visualise the three approaches.

In [19], the authors state that centralised solutions have enormously grown in the last years because of continuous streams of data being uploaded into cloud systems. Once collected, data can be used by ML-based systems for different purposes. Amazon, Google, and Microsoft are examples of ML-as-a-service providers in cloud systems. However, eavesdropping attacks pose privacy concerns for users. Given the risks of moving data to a centralised entity, on-site ML techniques have been proposed. The idea is to distribute a pre-trained/generic model to devices (clients) that, after deployment, can further refine (personalise) it using local data. Following such an approach, data never leave the host, thus mitigating risks and providing advantages over centralised solutions when it comes to privacy. Anyway, it must be noted that there is no benefit from other peers' data because local refinements are never shared. Additionally, if external algorithms (e.g., models) have to run on sensitive data, a secure processing environment (SPE) is needed to reduce risks coming from external code. To make other nodes (peers) benefit from local data while still preserving privacy, FL has been introduced. Also in the case of FL, SPEs are needed, but each node can also benefit from peers' model refinements. We proceed to a more thorough description of FL in Section IV.

IV. Federated Learning as Privacy-Enhancing Data Visiting Strategy

FL is a technique in which multiple parties (clients) collaborate to train a global model in federated settings, using local datasets that only designated clients can access and use. The task is carried out under the coordination of a central orchestrator. Local datasets are not shared among the clients and with the orchestrator, thus respecting data privacy. The process is depicted in Fig. 1, which shows the necessary components to be deployed (one orchestrator, N clients performing local training at N different data locations). A digital infrastructure, e.g., HPC, GPUs and other solutions, may be needed to run the learning processes, as elaborated in Section V. The costs associated with the use of FL are related to the achievable utility level and the increased complexity of the system [20].

FL proves of particular interest when information is stored in data islands, as common nowadays. Cloud computing and, more generally, centralised approaches, foresee data use in central locations into which data are collected, then processed and analysed. The collection of data in a central repository makes it a valuable target to attack, especially in the presence of sensitive data. Because of that, data providers may be unwilling to share data because of privacy concerns and fears of unauthorized access, misuse, or loss. If we put concerns and threats aside for a moment, several techniques can be used to extract value from data hosted in a central location. For instance, machine learning techniques can be used to train models. However, it is rather difficult for data providers to monitor access to their data and prevent or revoke it if necessary. Thus, compared with centralised training, FL enables data to be collaboratively used in controlled settings.

Some aspects must be emphasised to fully evaluate the pros and cons of FL. On the utility side, i.e., the accuracy of the global model, it must be noted that centralised settings may provide better results [20] but at the expense of data being shared. On the complexity side, the process of setting up and deploying an orchestrator and multiple clients has higher economic and technical costs, although frameworks for easy deployment and use are rapidly progressing. In Table I we provide an overview of available FL frameworks to show the increasing variety of products and services available today.

In Section V, the architecture we propose for using FL in data spaces is described, and in Section VI we better elaborate on the limitations and challenges of the approach we describe.

V. A Strategy for Implementation

As anticipated, there are already lots of frameworks implementing FL (see Table I) and proposals for infrastructures supporting data visiting, such as the one in [21]. Generally, such frameworks require the deployment of a client module at each Data Provider's (DP) infrastructure. Clients are controlled by an orchestrator in charge of coordinating the training tasks carried out by each client. Thus, all participating DPs must agree on the specific framework(s) to be used and maintain/update the client deployment on their infrastructure. This results in a tightly coupled architecture, which may be

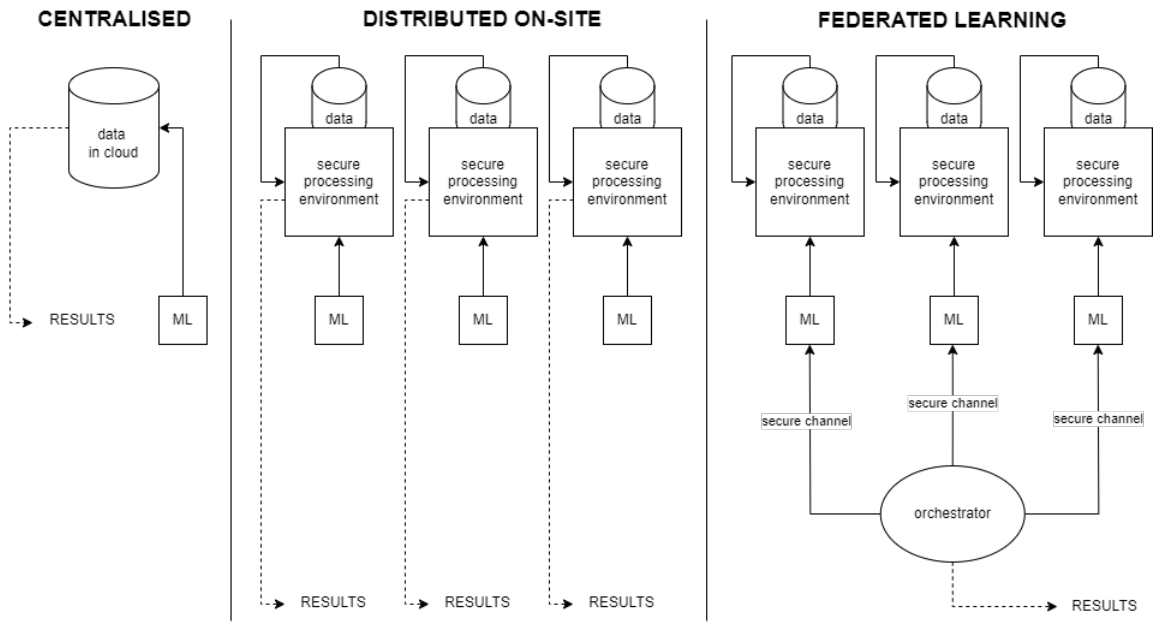

Fig. 2: Comparison of high-level architectures of centralised, distributed on-site, and FL solutions.

difficult to achieve and maintain in the current landscape of heterogeneous data-sharing systems and initiatives, including the Common European Data Spaces. This section proposes a high-level domain-agnostic architecture flexible enough to accommodate the varying needs of DPs and run different FL implementations on their infrastructure.

One of the key features needed to enable FL is the possibility to move code -e.g., training code implemented by an FL client- to the data (compute to data). This requires being able to utilize different computing infrastructures (e.g., cloud infrastructures, HPC, etc.) for the execution of computational processes. The architecture (technical blueprint) introduced in the Green Deal Data Space (GDDS) [22], here used as a reference approach, introduces a set of useful logical components. By acting as intermediaries, the GDDS architecture components implement the necessary functionalities to enable the GDDS Digital Ecosystem (DE). Among these components, the Data Processing Enabler (DPE) is specifically dedicated to enabling the execution of scientific models, including ML-based models, on different computing infrastructures. The DPE can be invoked by data consumers, which submit the execution request of their algorithm implementations, and it takes care of setting up the execution environment on the computing infrastructure, triggering the execution, and saving the output. A possible implementation of the DPE concept is the Virtual Earth Laboratory (VLab) [23] [24], an orchestrator framework enabling the execution of scientific environmental models in a multi-cloud environment. As said above, FL learning implementations must orchestrate the FL clients. To this aim, a DPE requires a set of enhancements specifically tailored to support one (or more) FL client-server framework(s).

We recall the NFI-related example introduced in Section II, which aims at enabling access to NFI data to calibrate geospatial machine learning models that underpin the development and delivery of forest indicators while maintaining the confidentiality of plot locations. Figure 3 depicts a UML component diagram showing how different components interact in the training phase of the NFI use case. The Forest Monitoring Downstream Service (FMDS) is an application developed to calculate forest indicators using an ML-based model. Each DP (i.e., NFIs) provides a Data Source (a service providing data discovery and access functionality) and a Computing Infrastructure (accessible via infrastructure-as-a-service/platform-as-a-service APIs). The FMDS queries the Data Catalog for forest data. After discovering the required data, it submits a FL request to the DPE. The latter relies on the DPs' Computing Infrastructures to execute the FL client application, thus locally accessing the data for training purposes without moving them. The Dataset Transformer component can be used by the DPE for any data transformation which might be required (e.g., format encoding).

The main advantages of the architectural design we propose are its flexibility and extensibility. The flexibility stems from the possibility of allowing the participation of different DPs without imposing any specific technological solution (i.e., FL framework); the only requirement for DPs is to provide a Computing Infrastructure that the DPE can use. Besides, the architecture can be extended to support new FL frameworks without the need to modify the interaction with either the Data Consumers or Providers.

VI. Limitations and Challenges

In this section, we briefly touch upon the limitations and challenges of FL and the strategy we propose. As antici-

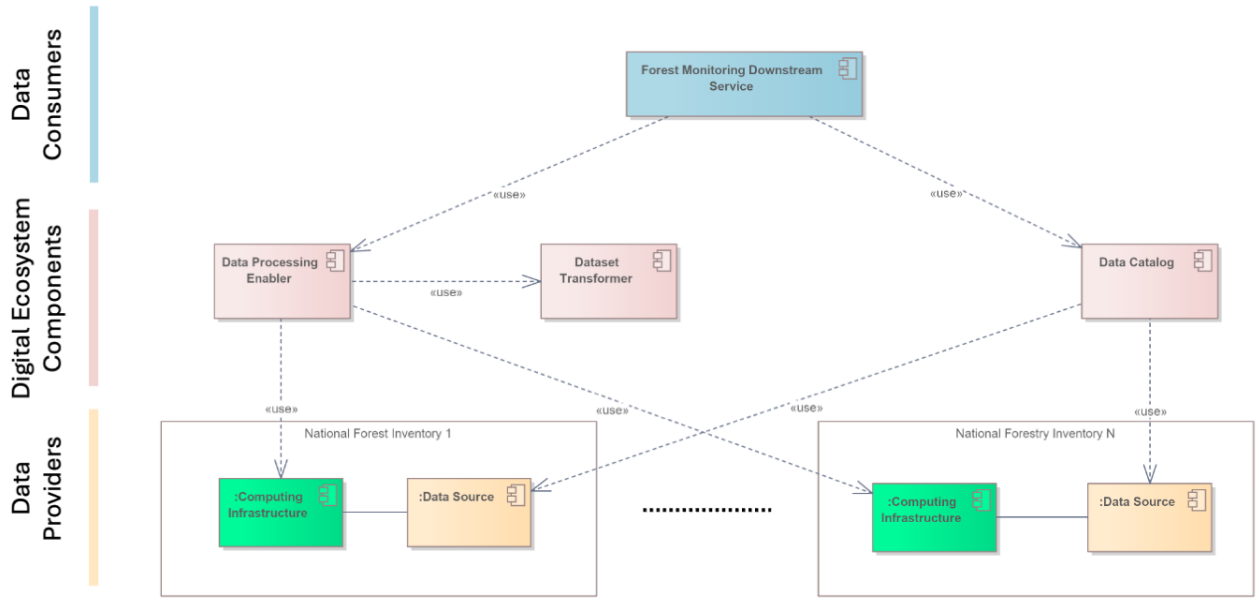

Fig. 3: UML diagram of the proposed architecture: the FMDS use data by several Data Providers (NFIs in our example) without moving it by executing FL instances on the DP (client-side) Computing Infrastructure under the DPE orchestration.

pated, the use of FL -and in general PETs- must take into account a lower accuracy (utility) if compared with centralised techniques; how much such a limitation could be critical depends on the specific domain and case. Another aspect to be considered is that setting up an infrastructure, like the one we propose, has development and maintenance costs to be considered; however, in several scenarios, a PET-enabled approach could represent the only viable option to use data, which otherwise would be inaccessible. The interoperability of the different FL client-server frameworks to be supported in our proposal, as cited in Section V, is an additional aspect to be carefully considered.

Explicitly considering FL, open challenges must be carefully considered when data distribution is uneven or non-IID (non-independent and identically distributed) across DPs. In fact, non-IID data can lead to significant differences in the data distributions across clients; this statistical heterogeneity makes it challenging for the global model to perform well on each client's specific task. The uneven data distribution can cause model drift, where the global model diverges from the optimal model for each client due to the biased gradients from clients with limited data. Furthermore, non-IID data can make it difficult for the federated learning algorithm to converge to a good global model, as the gradients from different clients may be conflicting [25], [26]. Adding to that, when data are unevenly distributed, it becomes challenging to detect class imbalances in the aggregated gradients, as the training data is not directly observable by the server [27].

Another class of challenges is interoperability at the semantic level. Consider again the the NFI data case we presented above: datasets from member states (MS) are collected following different protocols, using different units, different formats, different coordinate systems, etc. according to country-specific inventory designs. While this practice preserves comparability over the years in an MS, it makes it hard to compare data of different MS. Creating a pipeline for conversion and harmonization of data is no trivial task due to metadata being in different languages and of different quality, and even different definitions are applied by NFIs [28], [29]. This, and other classes of challenges we do not discuss in this work, may be addressed by data intermediaries as foreseen in the data space ecosystem, opening new market opportunities for them. Finally, we stress that trust and governance are to be considered in any case. Data visiting still requires agreements and trust relationships among participants, which takes time to build. The same goes for the definition of a commonly agreed governance mechanism.

VII. Conclusions

In this work, we explored data visiting as a strategy to use data otherwise inaccessible. This is in line with the idea of a data market, enabled by the use of PETs in the context of data spaces. We focussed on FL as the most relevant technique for the forestry use case we use as a reference, in which the sensitive data from the EU NFIs can be leveraged to train a global model to be used to support environmental efforts and the Green Deal initiative. We proposed a domain-agnostic architecture to run FL and explored the limitations and challenges for it to be increasingly effective in future.

Acknowledgement

Part of the research leading to the presented results has received funding from the European Union's Digital Europe

Programme under grant agreement N. 101083927 (GREAT).

DISCLAIMER

The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

REFERENCES

[1] European Parliament and Council of the European Union, "Regulation (eu) 2022/868 of the european parliament and of the council of 30 may 2022 on european data governance and amending regulation (eu) 2018/1724 (data governance act)," 2022. [Online]. Available: http://data.europa.eu/eli/reg/2022/868/oj

[2] ——, "Regulation (eu) 2023/2854 of the european parliament and of the council of 13 december 2023 on harmonised rules on fair access to and use of data and amending regulation (eu) 2017/2394 and directive (eu) 2020/1828 (data act)," 2023.

[3] "Digital (the digital europe programme)," https://digital-strategy.ec.europa.eu/en/activities/digital-programme, accessed: 06-27-2024.

[4] European Commission, "Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions. a european strategy for data (com/2020/66 final)," 2020.

[5] M. Weise and A. Rauber, "A Data-Visiting Infrastructure for Providing Access to Preserved Databases that Cannot be Shared or Made Publicly Accessible," in *iPRES*, 2021.

[6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016, https://doi.org/10.1038/sdata.2016.18.

[7] "Data Spaces Support Centre - Data Spaces Blueprint v1.0," 2024.

[8] D. H. Ramírez, L. P. Díaz, S. Rahimian, J. M. A. García, B. I. Peña, Y. Al-Khazraji, Á. J. G. Alarcón, P. G. Fuente, J. S. Garrido, and A. Kotsev, "Technological Enablers for Privacy Preserving Data Sharing and Analysis," 2023.

[9] European Commission and Joint Research Centre, E. Farrell, M. Minghini, A. Kotsev, J. Soler-Garrido, B. Tapsall, M. Micheli, M. Posada, S. Signorelli, A. Tartaro, J. Bernal, M. Vespe, M. Di Leo, B. Carballa-Smichowski, R. Smith, S. Schade, K. Pogorzelska, L. Gabrielli, and D. De Marchi, *European data spaces – Scientific insights into data sharing and utilisation at scale*. Publications Office of the European Union, 2023, https://data.europa.eu/doi/10.2760/400188.

[10] ENISA - European Union Agency for Cybersecurity, P. Drogkaris, and J. Gomez Prieto, "Engineering personal data protection in eu data spaces," 2024, https://www.enisa.europa.eu/publications/engineering-personal-data-protection-in-eu-data-spaces.

[11] R. S. Antunes, C. André da Costa, A. Küderle, I. A. Yari, and B. Eskofier, "Federated Learning for Healthcare: Systematic Review and Architecture Proposal," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 4, pp. 1–23, 2022.

[12] J. Wang and F. Ma, "Federated Learning for Rare Disease Detection: a Survey," *Rare Disease and Orphan Drugs Journal*, vol. 16, 2023.

[13] S. Pati, U. Baid, B. Edwards, M. Sheller, S.-H. Wang, G. A. Reina, P. Foley, A. Gruzdev, D. Karkada, C. Davatzikos *et al.*, "Federated Learning Enables Big Data for Rare Cancer Boundary Detection," *Nature communications*, vol. 13, no. 1, p. 7346, 2022.

[14] B. Chen, T. Chen, X. Zeng, W. Zhang, Q. Lu, Z. Hou, J. Zhou, and S. Helal, "DFML: Dynamic Federated Meta-Learning for Rare Disease Prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023.

[15] D. Li, W. Xie, Y. Li, and L. Fang, "FedFusion: Manifold Driven Federated Learning for Multi-Satellite and Multi-Modality Fusion," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[16] F. Luo, L. Liu, G. G. Wang, V. Kumar, M. S. Ashton, J. Abernethy, F. Afghah, M. H. M. Browning, D. Coyle, P. Dames *et al.*, "Artificial Intelligence for Climate Smart Forestry: A Forward Looking Vision," in *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2023, pp. 1–10.

[17] U. Hewage, R. Sinha, and M. A. Naeem, "Privacy-Preserving Data (Stream) Mining Techniques and Their Impact on Data Mining Accuracy: a Systematic Literature Review," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 10 427–10 464, 2023.

[18] J. P. S. Piest, W. Datema, D. R. Firdausy, and H. Bastiaansen, "Developing and Deploying Federated Learning Models in Data Spaces: Smart Truck Parking Reference Use Case," in *International Conference on Enterprise Design, Operations, and Computing*. Springer, 2023, pp. 39–59.

[19] S. AbdulRahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani, "A Survey on Federated Learning: the Journey from Centralized to Distributed On-Site Learning and Beyond," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5476–5497, 2020.

[20] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A Survey on Federated Learning," *Knowledge-Based Systems*, vol. 216, 2021.

[21] M. Weise, F. Kovacevic, N. Popper, and A. Rauber, "OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting," *Data Science Journal*, vol. 21, pp. 4–4, 2022.

[22] M. Santoro, P. Mazzetti, K. De Jong, C. Briese, M. Gutierrez, P. Gondim van Dongen, R. Oonk, G. Venekamp, and N. Raczko, "Final Blueprint of the GDDS Reference Architecture." [Online]. Available: https://www.greatproject.eu/wp-content/uploads/2024/04/D3.2-Final-Blueprint-of-the-GDDS-Reference-Architecture.pdf

[23] M. Santoro, P. Mazzetti, and S. Nativi, "The VLab Framework: An Orchestrator Component to Support Data to Knowledge Transition," *Remote Sensing*, vol. 12, no. 11, p. 1795, 2020.

[24] ——, "Virtual Earth Cloud: a Multi-Cloud Framework for Enabling Geosciences Digital Ecosystems," *International Journal of Digital Earth*, vol. 16, no. 1, pp. 43–65, 2023.

[25] V. N. Iyer, "A Review on Different Techniques Used to Combat the non-IID and Heterogeneous Nature of Data in FL," *arXiv preprint arXiv:2401.00809*, 2024.

[26] H. Lee, "Towards Convergence in Federated Learning via Non-IID Analysis in a Distributed Solar Energy Grid," *Electronics*, vol. 12, no. 7, 2023.

[27] L. Wang, S. Xu, X. Wang, and Q. Zhu, "Addressing Class Imbalance in Federated Learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 10 165–10 173.

[28] V. Avitabile and A. Camia, "An Assessment of Forest Biomass Maps in Europe Using Harmonized National Statistics and Inventory Plots," *Forest ecology and management*, vol. 409, pp. 489–498, 2018.

[29] C. Vidal, I. Alberdi, L. Hernández, and J. Redmond, "National Forest Inventories," *Springer Science+ Business Media. doi*, vol. 10, pp. 978–3, 2016.