# Personal Data Sovereignty through Federated Access and Policy Control

Vijon Baraku
Department of Computer Science, University of York
York, United Kingdom
vibaraku@seerc.org

Iraklis Paraskakis
SEERC - South East European Research Centre
Thessaloniki, Greece
iparaskakis@seerc.org

Simeon Veloudis
SEERC - South East European Research Centre
Thessaloniki, Greece
sveloudis@seerc.org

Poonam Yadav
Department of Computer Science, University of York
York, United Kingdom
poonam.yadav@york.ac.uk

## ABSTRACT

This paper presents a novel framework that enables individuals to exercise fine-grained control over their personal data across organisational boundaries. Despite regulatory advances like GDPR, individuals face significant challenges in maintaining sovereignty over their personal information due to data fragmentation across systems and lack of unified control mechanisms. Our solution addresses this through two key components: an Ontology-Based Data Federation system that creates a unified view of distributed personal data, and a Personal Data Policy Control service that enables governance. The framework allows individuals to discover their complete digital footprint and apply nuanced policies without requiring data migration or disrupting existing infrastructure. We present the implementation status, preliminary findings, and future research directions for enhancing personal data sovereignty.

**VLDB Workshop Reference Format:**
Vijon Baraku. Personal Data Sovereignty through Federated Access and Policy Control. VLDB 2025 Workshop: PhD Workshop.

## 1 INTRODUCTION

In today's quickly evolving digital society, data, data sharing, and the data economy serve as crucial enablers for digital transformation and are the lifeblood of modern economies and technological advancement [3, 9]. Within this data landscape, personal data stands out as a primary category of data being traded and processed. This data is highly valuable for businesses seeking to tailor services and gain consumer insights [5].

Despite the introduction of regulations such as the European Union's General Data Protection Regulation (GDPR), the Data Governance Act, and the Artificial Intelligence Act [1, 2] a critical problem persists: individuals lack granular control over their personal data. These frameworks, while establishing important principles, frequently fall short of empowering data generators, those who create data, to control how this data that refers to them is used. This limitation becomes particularly evident when organisations seek to use personal data for advanced processing scenarios [8].

The challenge is exacerbated by the increasing fragmentation of personal data across multiple systems. As personal information is distributed across numerous service providers, individuals struggle to maintain comprehensive awareness and control over their data footprint [4]. This fragmentation prevents individuals from exercising meaningful agency over how their information is collected, processed, and shared.

This research addresses the question: How can a framework enable individuals fine-grained control over their personal data while ensuring regulatory compliance? We propose a novel data sovereignty framework that not only ensures adherence to prevailing pertinent regulations but goes a step further by introducing a data capsule concept that aims at fundamentally redefining the way individuals control third-party access to their personal data.

The contribution of this research is a practical implementation that transforms theoretical data sovereignty concepts into executable systems capable of providing individuals with granular control over their distributed personal data, without disrupting existing data infrastructure.

The remainder of this paper is structured as follows: Section 2 presents the system overview and architecture. Sections 3 and 4 detail the ontology-based data federation system and personal data policy control service respectively. Section 5 concludes with preliminary results and future work.

## 2 RESEARCH APPROACH AND BACKGROUND

This section outlines the research approach and contextualises the work within relevant literature, highlighting limitations in current approaches and establishing the research objectives that guide this implementation.

### 2.1 Limitations of Current Approaches

Current approaches to personal data sovereignty fall into two primary architectural paradigms, each with significant limitations for achieving individual control over distributed personal data.

The first approach, exemplified by the Solid project [7], implements personal data stores or "pods" where individuals centralise their personal information under direct control. While providing clear ownership boundaries, this approach requires substantial


Proceedings of the VLDB Endowment. ISSN 2150-8097.

changes to existing data infrastructure. The primary limitation is the data migration requirement: Solid necessitates moving data from organisational systems into personal pods, creating significant adoption barriers. Additionally, pod-based architecture restricts discovery to data within the pod ecosystem, with no mechanisms for identifying personal data scattered across organisational systems.

The second approach, represented by the International Data Spaces Association (IDSA) framework [6], maintains data in its original location while establishing standardised protocols for secure exchange with usage control. While providing strong organisational sovereignty, IDSA primarily addresses business-to-business exchange rather than individual data sovereignty. Its connector-based architecture emphasises point-to-point data transfer rather than unified discovery across multiple sources. Furthermore, its usage control policies lack the specificity required for advanced governance scenarios such as AI training.

Neither approach adequately addresses the dual requirements of data visibility and governance necessary for personal data sovereignty in distributed environments. The Solid project focuses on control without addressing discovery across external systems, while IDSA enables controlled sharing without focusing on individual sovereignty rights.

## 2.2 Research Objectives

Addressing these limitations, this research defines four key objectives that guide the system implementation:

**RO1: Define mechanisms to enable data federation** - The first objective addresses the fundamental challenge of data heterogeneity and fragmentation. Given that personal data comes from numerous heterogeneous sources in different formats, federation mechanisms are essential for creating a comprehensive view. This research extends Ontology-Based Data Access (OBDA) to an ontology-based data federation scheme.

**RO2: Define mechanisms for efficient data annotation and classification** - The second objective focuses on enabling individuals to classify their data according to sensitivity, purpose, or value. This classification forms the foundation for governance decisions, allowing for contextually appropriate policy application based on data characteristics.

**RO3: Define data governance policies** - The third objective establishes mechanisms for expressing governance rules that automate enforcement based on classification. This research investigates the integration of the Open Digital Rights Language (ODRL) into policy frameworks, enabling machine-readable expressions of individual preferences.

**RO4: Guarantee trust** - The final objective addresses transparency, data provenance, and security. While noted as a research objective in the framework, this component primarily represents infrastructural considerations rather than novel research contributions.

## 2.3 Methodology

The research adopts a pragmatic approach, beginning with conceptual design and architecture and progressing through implementation and evaluation. The methodology followed these key phases:

(1) Understanding the motivation by identifying fundamental drivers and problems with current data handling approaches
(2) Conducting a comprehensive literature review to identify gaps in personal data sovereignty implementations
(3) Designing a framework architecture to achieve data sovereignty and meet the research objectives
(4) Implementing the system through a phased approach with iterative testing and refinement

This methodology ensures that the implementation addresses real-world challenges while building on established techniques in data management and governance. The next sections detail the resulting system components, beginning with the data federation layer that creates the foundation for sovereignty.

# 3 SYSTEM OVERVIEW AND ARCHITECTURE

The proposed data sovereignty framework implements a comprehensive approach to address the challenges of personal data discovery and control across organisational boundaries. This section presents the overall system architecture before examining individual components in subsequent sections.

## 3.1 Core Components

The framework comprises two primary components that work together to enable personal data sovereignty:

- **Ontology-Based Data Federation (OBDF) System:** Creates a unified view of personal data scattered across organisational databases by using Schema.org as a common vocabulary. This component addresses the fundamental challenge of data fragmentation without requiring data migration.
- **Personal Data Policy Control (PDPC) Service:** Enables individuals to define governance policies for their data using extensions to the Open Digital Rights Language (ODRL) standard. This component transforms visibility into actionable control through standardised policy expressions.

Supporting these core components is the Authentication Service that manages user identification and access control, ensuring appropriate boundaries while facilitating necessary interactions between stakeholders.

## 3.2 User Roles and Workflows

The system supports two primary user roles with distinct workflows:

(1) **Data Generators (Individuals):** Discover their personal data across connected controllers and define governance policies. Their workflow includes data discovery and policy management.
(2) **Data Controllers (Organisations):** Connect their databases to the system and map their schemas to the common vocabulary. Their workflow includes database integration and schema mapping.

These complementary workflows demonstrate how the architecture balances the interests and needs of both individuals and organisations, creating a framework for responsible personal data

management that respects individual sovereignty without disrupting organisational operations.

The following sections examine each core component in detail, beginning with the federation system that provides the foundation for data discovery.

# 4 ONTOLOGY-BASED DATA FEDERATION SYSTEM

The Ontology-Based Data Federation (OBDF) system implements virtual integration that maintains personal data in original repositories while providing subjects with a unified view. This component addresses the challenge of personal data fragmentation without requiring centralisation.

## 4.1 Federation Approach

The federation system uses Schema.org as a common vocabulary to create semantic bridges between diverse database schemas. This choice offers comprehensive coverage of personal data domains, established extension mechanisms, and leverages widespread industry adoption.

The architecture separates the concerns of database connection, schema mapping, query execution, and data presentation. Controllers connect their databases to the system and map their schemas to Schema.org concepts. These mappings are then translated into formal Ontology-Based Data Access (OBDA) definitions that the query engine uses to transform SPARQL queries into SQL for each data source.

## 4.2 Federation Workflow

The federation process follows a structured workflow that enables cross-organisational data discovery without centralisation. First, data controllers register their databases with the system and securely store connection parameters, establishing the foundation for federated access while maintaining organisational control. Controllers then use the mapping interface to connect their database elements to Schema.org concepts—for example, mapping patient records to the Schema.org Patient class and associated properties.

These mappings generate formal Ontology-Based Data Access (OBDA) definitions that enable query translation. When individuals request their personal data, the system executes parallel queries across all relevant databases, transforming a standard SPARQL query into database-specific SQL statements. The results are unified into a coherent view that maintains source information, allowing individuals to see which organisations hold their data while providing a comprehensive picture of their digital footprint.

This approach preserves regulatory compliance, avoids disruptive data migration, and scales efficiently with query complexity rather than data volume.

# 5 PERSONAL DATA POLICY CONTROL SERVICE

The Personal Data Policy Control (PDPC) service complements the federation layer by transforming data visibility into actionable sovereignty. This component enables individuals to define how their personal data should be processed through standardised policy expressions.

## 5.1 Policy Management Approach

The PDPC service adopts the W3C Open Digital Rights Language (ODRL) as its foundation for policy representation. ODRL provides a structured vocabulary for expressing permissions, prohibitions, and duties regarding digital content, making it well-suited for personal data governance. By leveraging this established standard, the system ensures that policies can be unambiguously interpreted across organisational boundaries.

A key design principle is the recognition that effective data sovereignty requires more than simple binary permissions. Real-world privacy preferences are nuanced and context-dependent, individuals may be comfortable sharing certain information for specific purposes but not others, or may place temporal limitations on data use. The policy model therefore incorporates multiple dimensions of control: access operations (reading, using, sharing, modifying), purpose limitations (service provision, research), temporal constraints (expiration dates), consequence mechanisms (notifications, compensation), and special-case restrictions (AI training, analytics).

## 5.2 Privacy Tier Templates

To balance expressiveness with usability, the system implements privacy tier templates that provide pre-configured policy sets appropriate for different sensitivity levels. These templates serve as starting points that individuals can customise based on specific preferences:

(1) **Tier 1: Public Data** - Minimal restrictions for non-sensitive information
(2) **Tier 2: Limited Sharing** - Internal processing only, prohibits external sharing
(3) **Tier 3: Sensitive Data** - Restricted usage with privacy-preserving requirements
(4) **Tier 4: Highly Restricted** - Minimal access for essential services only

Each tier configuration includes appropriate settings across all policy dimensions, from basic permissions to specialised controls for AI training and analytics.

## 5.3 ODRL Extensions for Advanced Governance

While the standard ODRL vocabulary provides a solid foundation, we have implemented extensions to address specific governance scenarios not covered by the core specification. These extensions include provisions for AI training governance, automated notification requirements, and compensation mechanisms.

For example, our AI training extensions enable individuals to specify constraints such as complete prohibition, algorithm-specific restrictions (federated learning, differential privacy), or purpose-limited usage. This granular control over machine learning processes represents a novel contribution to personal data sovereignty not present in existing frameworks.

# 6 PRELIMINARY RESULTS AND FUTURE WORK

## 6.1 Implementation Status

The current implementation demonstrates the viability of our approach to personal data sovereignty. The system architecture has moved from design to functional prototype, with core components already operational.

The Ontology-Based Data Federation system has been implemented with support for major relational database types. The Schema.org integration and OBDA mapping generation are fully functional, enabling cross-database query execution with appropriate performance for interactive use. Users can successfully discover their personal data across multiple organisations through a unified interface.

The Personal Data Policy Control service has implemented the basic ODRL policy model with custom extensions. The privacy tier templates are operational, allowing for simplified policy creation while maintaining expressiveness. Integration between the federation and policy components ensures consistent data element references across both systems.

Initial testing with synthetic datasets representing common personal data scenarios (healthcare, e-commerce, social platforms) shows promising results for both discovery and governance capabilities.

## 6.2 Current Limitations

Several limitations in the current implementation represent opportunities for improvement:

- **Policy Enforcement:** While policy expression is well-developed, technical enforcement mechanisms require further implementation, particularly for specialised scenarios such as AI training restrictions.
- **Scalability Testing:** The current prototype has been tested with moderate-scale datasets; comprehensive performance evaluation with large-scale deployments is still pending.
- **Trust Infrastructure:** The fourth research objective (RO4) regarding blockchain integration for transparency and non-repudiation remains at the design stage without full implementation.

## 6.3 Future Research Directions

Based on the current implementation status and identified limitations, several key directions for future work have been identified:

(1) **Policy Enforcement Mechanisms:** Developing robust enforcement of expressed policies through integration with privacy-preserving technologies like federated learning and differential privacy for AI training governance.
(2) **Blockchain Integration:** Implementing the trust layer to provide immutable policy records and transparent enforcement logs, creating an auditable trail of data usage activities.
(3) **Deployment Case Studies:** Partnering with organisations across sectors to evaluate the framework in production environments, gathering insights on real-world challenges and adoption considerations.

## 6.4 Research Contributions

This work makes several significant contributions to personal data sovereignty research:

- A practical implementation that transforms theoretical sovereignty concepts into executable systems
- A virtual federation approach that addresses data fragmentation without requiring migration
- ODRL extensions that enable nuanced policy expression for emerging governance scenarios

By enabling individuals to both discover and control their personal data across distributed environments, this research establishes a foundation for meaningful data sovereignty that respects both individual rights and organisational responsibilities.

# ACKNOWLEDGMENTS

This work is also supported, in part, by EPSRC and DSIT TMF-uplift: CHEDDAR: Communications Hub For Empowering Distributed ClouD Computing Applications And Research (EP/X040518/1), (EP/-Y037421/1), EPSRC IAA (EP/X525856/1), and EPSRC REMOTE (EP/-Y019229/1).

# REFERENCES

[1] European Commission. 2021. *Proposal for a regulation on European data governance - Data Governance Act.* Technical Report. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767
[2] European Parliament and of the Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L119 (2016), 1–88. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679
[3] Patrik Hummel, Matthias Braun, Max Tretter, and Peter Dabrock. 2021. Data sovereignty: A review. *Big Data & Society* 8, 1 (2021), 1–17. https://doi.org/10.1177/20539517211014542
[4] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. 2020. Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly* 37, 3 (2020), 101493. https://doi.org/10.1016/j.giq.2020.101493
[5] Brent D. Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 1–21. https://doi.org/10.1177/2053951716679679
[6] Boris Otto and Matthias Jarke. 2019. Designing a multi-sided data platform: findings from the International Data Spaces case. *Electronic Markets* 29, 4 (2019), 561–580. https://doi.org/10.1007/s12525-019-00362-x
[7] Andrei V. Sambra, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, Abdurrahman Ghanem, Dimitri Zagidulin, Ashraf Aboulnaga, and Tim Berners-Lee. 2016. Solid: A Platform for Decentralized Social Applications Based on Linked Data. *Technical Report, MIT CSAIL & Qatar Computing Research Institute* (2016). https://solid.mit.edu/
[8] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2017), 841–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf
[9] World Economic Forum. 2020. *Data-driven economies: Foundations, frameworks, and ethical perspectives.* Technical Report. World Economic Forum. https://www3.weforum.org/docs/WEF_Data_Driven_Economies_2020.pdf