Toward Trustworthy Personal AI: A Sovereign Agentic Web Architecture Using Solid, Federated Learning, and Personal Knowledge Graphs

Fernando Spadea, Lorenzo Carta, Abhirup Dasgupta, Md Saikat Islam Khan Bappy and Oshani Seneviratne

Abstract

The emergence of autonomous Personal Artificial Intelligence (AI) requires profound contextual understanding of user data, creating a direct tension between algorithmic utility and data sovereignty. While decentralized storage infrastructures like Solid decouple data from applications, they traditionally treat personal data pods as passive repositories rather than substrates for active intelligence. In this paper, we summarize our recent contributions toward building a "Sovereign Agentic Web." We present a synthesis of a decentralized architecture that fuses Solid pods with Federated Learning and Differential Privacy to achieve privacy-preserving, user-centric AI. By anchoring Large Language Models to Solid pods, utilizing Personal Knowledge Graphs for semantic grounding, and establishing rigorous protocol adaptations for agent-to-agent communication, we demonstrate that rigorous privacy guarantees can coexist with the predictive utility required for trustworthy, socially responsible AI agents.

Keywords

Solid, Federated Learning, Agentic Web, Data Sovereignty, Edge AI, WebGPU, Privacy-Preserving ML, Decentralized AI

1. Introduction

The centralization of AI training pipelines has established a paradigm where sensitive user data is ingested into proprietary silos, fundamentally eroding data sovereignty and user trust. While Federated Learning (FL) decentralizes computation by bringing the model to the data [1], it does not inherently solve the issue of long-term data ownership or user-centric governance. Concurrently, the Solid protocol empowers users to retain control over their data via Personal Online Datastores (pods) [2], yet current implementations largely treat these pods as passive storage without taking full advantage of user control. Existing personalization architectures also face semantic fragmentation, which prevents AI systems from reasoning over heterogeneous personal data. There is also a lack of governance mechanisms for autonomous agents, which creates risks around delegation, accountability, and consent.

To meet the goals of trustworthy autonomous personal AI, we argue that the optimal path forward lies in the fusion of these technologies. In our recent work, we proposed a comprehensive reference architecture for a "Sovereign Agentic Web" on the Solid ecosystem [3, 4]. This framework anchors FL clients directly to Solid pods to ensure strictly local model personalization, establishes protocol-centric communication patterns to interact with pods and other agents through authenticated, policy-aware exchanges of semantically structured messages, and leverages semantics to overcome data heterogeneity. This paper synthesizes these contributions, outlining the data structures, computational paradigms, and protocol adaptations necessary for personal AI.

Concretely, our combined architecture integrates four aspects: (1) pod-resident Personal Knowledge Graphs (PKGs) that store semantically structured user data, (2) local LLM-based reasoning agents operating directly on the pod, (3) FL mechanisms that enable collaborative model improvement without

ESWC'26: Trust, Autonomy and Accountability in PKG-Based Agentic AI (TAAPAAI) Workshop on May 10, 2026 in Dubrovnik, Croatia

spadef@rpi.edu (F. Spadea); cartal@rpi.edu (L. Carta); dasgua3@rpi.edu (A. Dasgupta); islamm9@rpi.edu (M. S. I. K. Bappy); senevo@rpi.edu (O. Seneviratne)

0009-0006-4278-3666 (F. Spadea); 0009-0005-5610-9093 (L. Carta); 0009-0002-8434-9754 (A. Dasgupta); 0009-0009-1768-6102 (M. S. I. K. Bappy); 0000-0001-8518-917X (O. Seneviratne)

published 2026-06-09

centralizing data, and (4) protocol-level adaptations that enable safe coordination between autonomous agents across pods.

2. Solid as an Active Data Substrate for Personal AI

Personalized AI requires deep, context-aware user data and models. Traditional recommendation systems and autonomous agents rely on centralizing raw interaction data. We invert this model by utilizing Solid pods as the secure Information Retrieval (IR) substrate for PKGs [5]. In our architecture, PKGs provide structured, semantically grounded context, while LLMs perform reasoning and natural language interaction over this structured context.

2.1. Semantic Retrieval of Evolving PKGs

By hosting PKGs natively as RDF documents within Solid pods, we formulate the personal AI recommendation as a decentralized IR task. When a user issues a query (typically in natural language), it is semantically analyzed to identify the target domain. This populates a parameterized SPARQL template that executes against the Solid pod where the PKG is hosted, retrieving only task-relevant triples (e.g., isolating financial data from health records). A lightweight LLM operating locally on the user's device performs the semantic interpretation and recommendation reasoning, while FL enables these models to improve collaboratively across users without requiring direct access to their personal data.

To accommodate the strict token limits of lightweight edge LLMs, we apply a semantic ranking step. Triples are ranked based on temporal recency or semantic similarity to the user's query, ensuring only the most contextually vital nodes are injected into the LLM's prompt.

As users interact with recommendations, their feedback (e.g., accepting, rejecting, or modifying suggestions) is translated into new semantic triples that capture inferred preferences, contextual signals, and interaction outcomes. These triples are committed back to the Solid pod, creating a *continually evolving PKG* [5]. Over time, this process enriches the PKG with both observed and inferred knowledge, enabling more accurate personalization while ensuring that the resulting knowledge remains locally governed and under the user's control.

3. Privacy-Preserving Federated Intelligence

To achieve collective intelligence without data exfiltration, we rely on FL. We demonstrated the viability of this approach through a high-stakes use case: predicting consumer financial distress using data derived from the U.S. National Financial Capability Study (NFCS). In our experiments, individual survey responses are treated as distributed data points residing in separate simulated pods, allowing the model to be trained through federated aggregation while preserving the locality of personal financial information simulated to be individual data points [6].

3.1. Federated Analytics and Differential Privacy

In our financial risk modeling application [6], the FL client acts as an authorized agent executing within a user-controlled trusted sandbox environment adjacent to the user's pod. By co-locating model training with the pod, raw personal data never leaves the user's storage boundary. To mitigate inference risks associated with shared model updates we integrate Differential Privacy (DP) [7] directly into the local training pipeline. During training. model updates are clipped to a fixed L_2 norm to bound sensitivity, and statistically calibrated Gaussian noise is injected before aggregation. This mechanism ensures the resulting updates satisfy (ϵ, δ)-DP guarantees, providing provable protection against reconstruction attacks while preserving the aggregate utility needed for early warning systems [8].

3.2. Web-Native Edge Compute

Executing high-dimensional model updates on consumer-grade hardware or within browser runtimes presents a significant runtime bottleneck. Our architecture addresses this by utilizing WebGPU-accelerated Quantized Low-Rank Adaptation (QLoRA) [3]. The client hosts a frozen, quantized base LLM and only updates a small set of adapters. By offloading these parallel matrix operations to the local GPU via standard Web interfaces, we transform the browser into a viable edge-computing environment capable of sophisticated training tasks.

4. Protocol Adaptations for Sovereign Agents

While decentralized infrastructures emphasize user control, current agent ecosystems lack standardized, Web-native communication protocols. We propose specific protocol adaptations to elevate agents to first-class Web participants [4].

4.1. Agents as Web Principals

To enable decentralized coordination, agents are modeled as distinct Web Principals, i.e., entities with their own resolvable WebID identity, authentication credentials, and authorization context. In this context, an agent is treated as an independent actor on the Web, similar to a user or application, capable of authenticating, accessing resources, and interacting with other agents under explicitly delegated permissions. This allows users to delegate explicit, scoped, and revocable capabilities through standard Web-based identity and authorization mechanisms.

To streamline agent authentication, which can be computationally expensive and unintuitive via standard Solid OIDC [9], we propose the integration of Schnorr signatures [10] and X.509 certificates. By signing model updates and messages with a key pair linked to their WebID, agents provide non-repudiation without relying on a centralized directory [3]. This identity model enables agents to participate as first-class actors, allowing their actions, such as data access, model training, or coordination with other agents, to be cryptographically attributable and auditable. Because each agent maintains an independent WebID and key pair, delegation can be granular and revocable, ensuring that users retain ultimate control over which agents may operate on their pods and under what conditions.

4.2. Pod-to-Agent and Agent-to-Agent Coordination

We define three primary communication paradigms to ensure interoperability via safe and policy-compliant interactions between agents and personal data pods [4]. These paradigms are designed to ensure that agents can coordinate while respecting access control policies, minimizing unnecessary data exposure, and maintaining auditability of actions.

- **Pod-to-Agent:** We utilize the Model Context Protocol (MCP) [11] as a neuro-symbolic mediation layer. In this paradigm, the Agent (MCP Host) interprets natural language intents and invokes specific MCP Tools exposed by the pod-resident server. The MCP Server acts as a Policy Enforcement Point (PEP): it receives these structured requests and utilizes a specialized natural language processing service (or a semantic mapping engine) to transform them into formal queries (e.g., SPARQL) against the PKG. This ensures that only ontology-aligned, typed representations (e.g., `fhir:Patient`) are returned, contingent upon a successful policy check against the Solid Access Control Policies (ACPs) [12].
- **Agent-to-Agent:** Agents coordinate via an extended Solid Notifications Protocol [13]. Negotiations utilize semantically typed messages (e.g., `as:Offer`, `as:Accept`) posted to peer inboxes. This preserves information asymmetry; external agents learn only the outcome of a decision (e.g., a time slot is booked) rather than the private data motivating it (e.g., a medical diagnosis).

- **Policy-Carrying Data:** To move beyond rigid Access Control Lists, we advocate for Policy-Carrying Data, where flexible usage rules (e.g., ODRL [14]) are stored alongside the data. This allows agents to dynamically reason over temporal constraints or purpose-specific limitations at runtime.

4.3. Human-Centered Consent Lifecycle

Meaningful AI acceptability requires transparent governance. Our framework embeds consent enforcement directly into the system architecture [8]. Consent artifacts are treated as versioned, machine-interpretable objects that specify the authorized agent, permitted resources, and intended purpose. These consent policies are stored and enforced at the pod level, enabling fine-grained control over which agents may access or process personal data.

For example, a user may authorize a financial assistant agent to access selected financial triples in their PKG to provide budgeting recommendations or participate in a federated financial risk prediction model. The consent artifact would specify the authorized agent, the permitted data categories (e.g., income or credit history), and the intended purpose (e.g., financial risk analysis). If the user later withdraws consent, the associated access policies are updated at the pod level, immediately preventing the agent from further accessing the protected data.

To reduce cognitive load, pod-resident transparency dashboards dynamically translate these RDF-based policies into natural language explanations, allowing users to easily understand how their data is being used. This interface also enables users to monitor audit logs, DP budgets, and participation history in federated model training, enabling users to maintain awareness of how their data contributes to collective intelligence while retaining meaningful control over their personal information.

When coordinating across multiple agents and institutions, policy conflicts inevitably arise. For example, institutional requirements may intersect with user-defined preferences. To resolve these conflicts, our framework applies strict, conservative precedence rules. The system operates on a deny-by-default basis, and user-defined policies strictly dominate institutional defaults. In scenarios where conflicts persist, agents can propose less-invasive alternatives, such as reduced feature sets or stricter privacy budgets. This ensures that the user's data sovereignty is never compromised by external requirements or conflicting trusted sources.

4.4. Motivating Use Case: Multi-Agent Coordination

To ground this architecture, consider a cross-domain coordination scenario involving a council of specialized agents anchored to a user's Solid pod. A Health Agent monitors a user's biometrics. Using personal KGs, it maps sleep trends to medical ontologies, detects signs of burnout, and signals a need for recovery. Rather than merely alerting the user, it triggers a Scheduling Agent to block time. This agent uses local KGs to prioritize existing commitments and leverages FL-derived models to negotiate socially normative cancellations without exposing the user's private health status. This demonstrates how specialized agents can orchestrate complex tasks over private data; a capability unattainable by siloed cloud services.

5. Empirical Validation

While the integration of Solid and FL provides strong theoretical privacy guarantees, the practical efficacy of our foundational federated learning frameworks has already been demonstrated across multiple domains. These preliminary evaluations were conducted prior to full Solid integration; they serve as the predictive baseline for our architecture. In the context of consumer financial distress prediction using the U.S. National Financial Capability Study, our cross-silo FL approach achieved predictive performance nearly identical to centralized models [6]. Specifically, the federated model achieved a 42.2% F1 score and a 71.4% AUC, closely matching the centralized model's 42.4% F1 score and

74.1% AUC. Crucially, this federated approach vastly outperformed purely localized state-level models, which only achieved an average F1 score of 33.31%.

Furthermore, for personalized recommendations, our Federated Targeted Recommendations with Evolving Knowledge graphs and Language Models (FedTREK-LM) framework demonstrated substantial performance gains [15]. By locally fine-tuning lightweight LLMs (such as Qwen3 models) using Kahneman-Tversky Optimization, the federated framework achieved more than a 4x improvement in F1-score on recommendation benchmarks compared to state-of-the-art baselines like KBGAT and HAKE. These findings confirm the robustness of the decentralized learning paradigms that will power the Solid-based infrastructure.

6. Conclusion

Current AI ecosystems largely rely on centralized data aggregation, creating fundamental tensions between personalization, privacy, and user autonomy. In contrast, the architecture outlined in this paper demonstrates how these tensions can be addressed by grounding intelligent services directly within user-controlled data environments. The convergence of the Solid protocol, FL, and semantic technologies thus provides a robust foundation for a trustworthy and autonomous personal AI ecosystem.

By synthesizing our recent work on evolving PKGs, privacy-preserving financial analytics, web-native compute, and agentic protocol adaptations, we outline a feasible path toward a *Sovereign Agentic Web*. In this paradigm, personalization is achieved locally through semantic grounding, collective intelligence is derived via federated aggregation, and inter-agent coordination is governed by explicit, verifiable Web protocols.

More broadly, this approach reimagines the Web as a platform for accountable and user-aligned autonomous agents, where intelligence is decentralized, governance is transparent, and individuals retain meaningful control over their data and digital representatives. Future work will focus on the deployment of these mechanisms in live, cross-domain real-world testbeds and further refining the neuro-symbolic bridges that connect probabilistic learning systems with policy-aware, semantically grounded data governance.

Resources

Implementation artifacts are available at https://github.com/brains-group/SolidFL-Agents.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly for grammar and spelling checks, as well as for paraphrasing and rewording. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.

[2] A. V. Sambra, E. Mansour, S. Hawke, M. Zereba, N. Greco, A. Ghanem, D. Zagidulin, A. Aboulnaga, T. Berners-Lee, Solid: a platform for decentralized social applications based on linked data, MIT CSAIL & Qatar Computing Research Institute, Tech. Rep. 2016 (2016).

[3] F. Spadea, L. Carta, A. Dasgupta, M. S. I. K. Bappy, O. Seneviratne, Towards a sovereign agentic web: Fusing solid and federated learning for user-centric ai, 2026. Accepted to the 4th Solid Symposium Poster Session.

[4] O. Seneviratne, L. Carta, F. Spadea, Protocol adaptations for a sovereign agentic web in solid ecosystems, 2026. Under submission to Solid4Future Protocols session at the 4th Solid Symposium.

[5] F. Spadea, L. Carta, O. Seneviratne, Semantic retrieval of evolving personal knowledge graphs for federated llm recommendations via solid, 2026. Under submission to the Decentralised IR and AI based on Solid session at the 4th Solid Symposium.
[6] L. Carta, F. Spadea, O. Seneviratne, Explainable federated learning for us state-level financial distress modeling, arXiv preprint arXiv:2511.08588 (2025).
[7] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy, Foundations and trends® in theoretical computer science 9 (2014) 211–407.
[8] O. Seneviratne, F. Spadea, L. Carta, Designing privacy-preserving financial risk analytics on solid pods, 2026. Accepted to the 4th Privacy & Personal Data Management Session at the 4th Solid Symposium.
[9] A. Coburn, e. Pavlik, D. Zagidulin, Solid-oidc, Solid Project Technical Report, 2022. URL: https://solidproject.org/TR/oidc.
[10] C. P. Schnorr, Efficient identification and digital signatures, Journal of Cryptology 4 (1991) 161–174.
[11] Anthropic, Model context protocol specification, https://github.com/modelcontextprotocol/specification, 2024. Version 1.0.
[12] M. Bosquet, Access Control Policy (ACP), Editor's Draft, Solid Community Group, 2022. URL: https://solid.github.io/authorization-panel/acp-specification/.
[13] Solid Community Group, Solid Notifications Protocol, Report, W3C Solid Community Group, 2022. URL: https://solidproject.org/TR/notifications-protocol.
[14] M. De Vos, S. Kirrane, J. Padget, K. Satoh, Odrl policy modelling and compliance checking, in: International Joint Conference on Rules and Reasoning, Springer, 2019, pp. 36–51.
[15] F. Spadea, O. Seneviratne, Federated personal knowledge graph completion with lightweight large language models for personalized recommendations, in: Proceedings of the European Semantic Web Conference, 2026.