

Introduce an Auditing Layer to Web Science

Rui Zhao
rui.zhao@cs.ox.ac.uk
University of Oxford
Oxford, England, UK

Jesse Wright
jesse.wright@jesus.ox.ac.uk
University of Oxford
Oxford, England, UK

Abstract

Scientific discoveries increasingly depend on data and data processing, and Web Science is no exception. As an established practice, data-intensive research typically uses scientific workflows and provenance to facilitate data and method sharing while automatically preserving processing history. Prior research has reported the possibility of ex-post policy-based compliance checking from provenance data. Based on these works, in this paper, we present the conceptual design of a framework of data-harvesting Web Science practices, especially by introducing a common auditing layer. We discuss the framework's practical, scientific, and ethical advantages, including its applicability in the period of large language model (LLM), autonomous agent, and artificial intelligence (AI) explosion. We hope this framework design can incubate a new norm for research practice to be transparent, ethical, and lightweight.

CCS Concepts

• **Information systems** → **Semantic web description languages**;
• **Security and privacy** → *Human and societal aspects of security and privacy*; • **Computing methodologies** → Cooperation and coordination.

Keywords

Data governance, transparency, audit, usage control

ACM Reference Format:

Rui Zhao and Jesse Wright. 2025. Introduce an Auditing Layer to Web Science. In *Proceedings of the 17th ACM Web Science Conference Companion 2025 (WebSci Companion '25)*, May 20–24, 2025, New Brunswick, NJ, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3720554.3736184>

1 Introduction

Many types of research nowadays depend on large volumes of data. Web Science is a particular example where many subfields, such as online behaviour or social media analysis like [15], depend on the harvesting of Web data and further processing. However, the willingness of the data providers is often overlooked, resulting in either complete ignorance or high practical burdens – for anyone wishing to assign any rules to their data, closing down access behind an application-and-review process is the only practical option. It further complicates the situation when we consider research reproducibility and open science [4], due to insufficient data access

to evaluate the results or unexpected non-compliance in sharing the data, let alone running experiments.

On the other hand, e-Science, or data-intensive research [5], a field also heavily depending on data processing, has an established practice of using the “scientific workflows” [1], forming a basis for FAIR research methods [20]. Such workflows¹ are composed of reusable units of processes, and can be applied to different data sources without modifications by connecting to a different data source at execution by workflow management systems (WMS). Further, many WMS possess the ability to generate provenance during execution, which fosters auditability. Despite the differences in WMS details, provenance is used to maintain interoperability, thanks to W3C PROV-O [16] and semantic technologies such as OWL [17] for reasoning. Its similarity to subfields of Web Science attracted some attempt in bridging the methods (e.g. [9]).

In the meantime, data are often subject to different governance rules or usage policies. Historical research such as Policy Aware Web [18] envisions describing usage control on web content with accountability mechanisms. More recent research has proposed different access control and usage control mechanisms to assist with this in different scenarios. In particular, Rui Zhao et al. [11] designed Dr.Aid for modelling and reasoning about such policies, using provenance from scientific workflows as input for retrospective analysis. Further research, such as psDToU [24], also explored the expressivity aspect of the policy language. But such research is often theoretical, and does not address an important question: where to obtain the information and policies for the reasoning?

In this paper, we describe a comprehensive framework design for Web Science practice, taking the lessons learned from e-Science and policy research. The framework focuses on interoperability, reusability, and auditability, addressing some common issues in existing research [6]. Further, this mechanism can be extended to be used by both human and software agents, where information exchange and gathering becomes more auditable, transparent, and reliable. We anticipate this framework to foster the discussion of more trustworthy and reliable research.

2 Design of framework for auditable Web Science

Here, we present the framework design as the basis of a new envisioned norm in Fig. 1. The framework consists of several components, serving different purposes, and eventually complement and cooperate with each other. It is worth noting that the services described here are not meant to be a “singleton”, but rather a federation of services of each type. Relevant mechanisms should be in place where the authenticity of information is required, as discussed later. With that in mind, the key components are introduced.

¹Not to be confused with the casual meaning of (human) actions and tools. We always refer to scientific workflows in this paper.



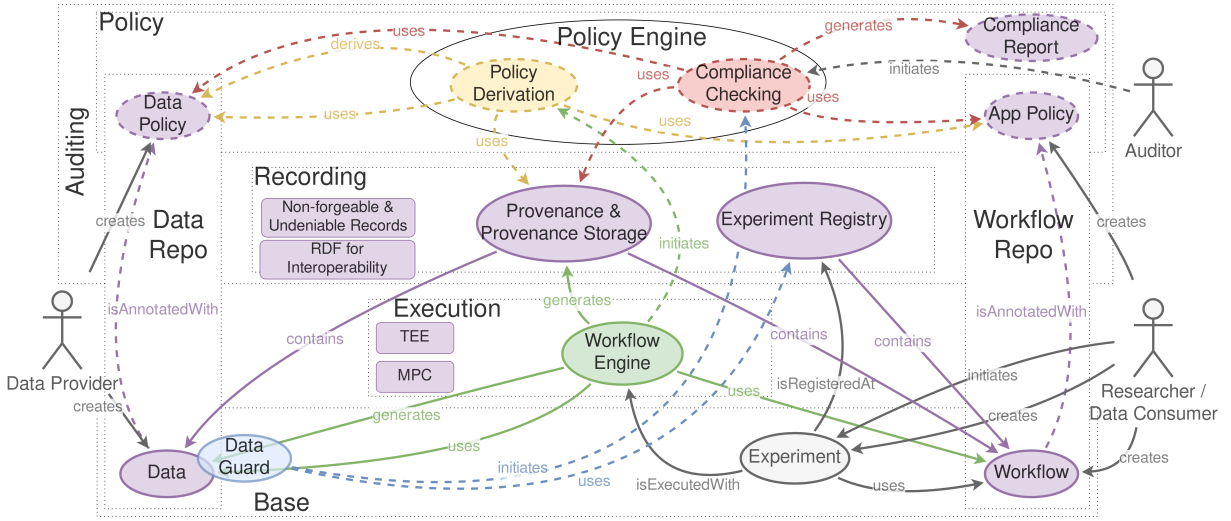


Figure 1: Components and their relations for the framework. Colors indicate different components/engines and the respective actions. Purple indicates elements that perpetually exist. Dashed lines indicate policy-related components and relations.

Workflow engine. An engine that deploys and executes a workflow, resembling a particular experiment to be performed. The workflow engine also records provenance of the execution trails.

Experiment registry. A service that holds the records of experiments, containing the workflow, the input source, and execution/deployment strategy. Other corresponding metadata, such as the operator and the corresponding policies of the workflow, will also be recorded. Unforgeable and undeniable records may be needed when logging the information to ensure the authenticity of information, as a means to establish or (im)prove trust.

Workflow repo and data repo. As an optional addition to foster scientific communication and reusing of existing methods, relevant repositories of workflows and data may be set up. This would also simplify the process of app policy authoring, as existing policies can be automatically fetched and reused if a new experiment reuses existing workflows or steps.

Provenance storage. Provenance is typically produced by the workflow engine, and then stored at a provenance storage. It is expected that no modification of provenance will be performed, due to the nature of such information. Appropriate measures for authenticity are also needed.

Policy engine. Policy engine has two main types of tasks to perform: compliance checking and policy derivation. Compliance checking is performed to verify if a data processing (e.g. a workflow or a relevant request) is compliant with the policies specified for input data. Policy derivation is used to produce policy for output data from a data processing, as a critical prerequisite to allow compliance checking to happen for downstream tasks.

Data guard. If policy-compliance of data processing is required, a guard shall be placed at every data source, before the data can be retrieved for processing by the workflow engine. The guard verifies the legitimacy of the request from the experiment registry,

performs policy compliance checking, and then releases the data for the workflow engine.

3 Usage and benefits of the framework

3.1 For data providers

The most prominent aim and benefit that this framework provides is the respect of data providers. Unlike normal data usage across the Web, heavy-weight data usage under this framework is governed by the policies assigned by the data providers. Data processors (e.g. researchers) need to first justify their intention before using the data, by the policy checking step of the data guard.

The basic form of the framework assumes that every actor will not deviate from the protocols. This is similar to current real-life practices, allowing easy transformation from current practice to the new framework without a heavy burden, and should generally be satisfactory for researchers in research institutions. If further guarantees are needed, appropriate authenticity mechanisms can be set up for relevant steps, as discussed later. However, such authenticity mechanisms will reduce the openness of Web Sciences. Further research is needed to develop better mechanisms without or with lower requirements for such closeness.

3.2 For respectful data consumers

Another side of the coin for respecting data providers is the benefits to the data consumers, typically researchers. Under contemporary practice, researchers are often subject to ethics approvals or other types of monitoring, or challenges to prove they conduct research ethically. However, such procedures are often time-consuming and non-portable – the fact that researcher B is doing identical research and received identical training as researcher A does not automatically grant B the clearance to conduct that research, nor does it reduce the burden B has to overcome.

With the new framework, researchers can use standard ways to describe a) the procedure they will follow to conduct the research, as

workflows, and b) the cautions they will take during the process, as app policies. Ethics committees can use this information and make a decision faster with the help of the automated policy reasoner. Similarly, this can also simplify the persuading procedure to data providers, who may have a different ethics committee or clearance team, making their own judgments.

In addition, for external auditors (including data providers), the records kept in experiment registry can be a source of verifying the genuineness of the experiment and the attitude of the data consumers. The data consumers can also use this as proof of their reputation, as well as a source for communication.

3.3 For LLM research and agents

Due to the wide and uncontrolled utility of scraping of Web content for training AI (artificial intelligence) models, many websites employ stricter rules for bots. This both reflects the insufficiency of traditional methods such as robots.txt, and the challenges of content ownership or, more generally, policies. In fact, many websites do not entirely prohibit using their data for LLM (large-language model) training, but want the data to be used in a more controlled manner.

The same fact causes issues with agents, as they are simply identified as bots and therefore denied access.

We anticipate the proposed framework can improve the situation, by providing better accountability and therefore trust. The case for LLM scrapers is much like that for other researchers, as illustrated above, where mutually agreeable policies would permit data usages to happen more naturally. For agents, the situation becomes more complicated due to their more diverse interactions.

However, despite the complexity, we can still witness a similar pattern: an agent has a stable identity with known capability, and interaction history can be recorded. This can fit into the framework where the experiment registry stores the agent's general information and capability (as well as policies), and its interaction history is recorded and stored in the provenance storage.

When it comes to agent-agent interactions, South et al. [14] identifies gaps in authorisation, accountability and access control capabilities of current LLM-based agents. Whilst standards such as OIDC4VP are readily deployable to delegate permissions, these do not provide effective mechanisms for describing the *actions* agents can perform and data they can use. Wright et al. [22] presents two personal agents conversing with one another on behalf of users; and illustrates how policies can be used to describe and sign data usage agreements - providing an audit trail to support accountability. Additionally, Wright et al. signs data and provenance sent between agents to support trust in data integrity and quality. Standards and regulations around Verifiable Credentials (VCs) [21] are important here to establish which information sent between agents issued by (1) real entities (e.g. humans and organisations) and (2) certified issuers of authoritative documents (e.g. Drivers License Authority).

3.4 Auditable & Reproducible

Many approaches nowadays require certain forms of auditability, which can be easily achieved in this framework, in a clean and reusable manner. Similar to practices like Open Science Framework

(OSF)², the experiment registry keeps records of experiments being registered (and maybe their ethics clearance status as well), including the personnel, method, and app policies. The provenance storage maintains the history of experiment execution, for further examination of whether the experiment has been conducted correctly; it also keeps records of data lineage, providing an auditable trail of data derivation and policy derivation, as well as a possibility for revision of data usage policies. Following lessons from e-Science, all records will be in linked data, to foster interoperability and long-term sustainability. This achieves FAIR [19] not only for data, but also for experiments and their executions, as they are resources as well. Likewise, with the recorded information in the registries and repositories (and appropriate access permission), reproducibility can be easily achieved by re-executing the same experiment by another researcher.

3.5 Generic

The framework does not assume a single operational entity for the services. Instead, all services can be a federation of the same type, as long as they follow the same protocol and maintain interoperability of records by following a common ontology. This avoids a single point of failure, not just for service stability, but also for reputation and maintainability. Expectedly, a research institution (or a group of collaborating research institutions) can maintain an experiment registry, provenance storage, and data guard, and provide these to researchers belonging to the institution.

In addition, this framework is useful not just for Web Science, but also for other types of research, e-Science included. They may not require the same level of all features (e.g. most e-Science may not require ethics clearance), but it will not impact the necessity of the rest of the system – they still need to prove to data providers of their reputation and respect for using data for their research, and sharing experiments with others.

3.6 Privacy and confidentiality

The word “privacy” has many connotations, where *confidentiality* and (*ability to*) *control* are two common examples. With the introduction of policy-related elements and persistent records, this framework provides a comprehensive solution for the *control* aspect, where data providers are able to exercise it by assigning stricter or looser policies to be used by data consumers. This allows more formal middle-grounds apart from normal practices, where either data is open to all, or protected with complicated requirements.

Confidentiality can be achieved by the cooperation between the policy engine and the workflow engine – the policy engine may require data processing to happen in a trusted non-leaky environment, and the workflow engine respects this during the execution. Privacy-preserving computation may also be employed for stronger guarantees, as discussed in Sec 4.2.

This framework is of more relevance if the execution happens inside a secure enclave, as a) data consumers need a more structured and complete description of their experiments before running tasks in a secure enclave, and b) no auditor can investigate details of the intermediate data or steps before concluding whether the final result can be released. Policy reasoning will be a main source of

²OSF: <https://osf.io> (accessed on 2025-04-30)

information to quickly screen the result’s appropriateness, before or after the execution.

4 Technical viability

Many components can be constructed using commodity technologies, including the workflow engine, experiment registry, repos, provenance storage and data guard. In the meantime, specific aspects worth separate discussion regarding the technology to enable the design.

4.1 Provenance storage and experiment registry

Apart from general technical implementations as a service, two additional considerations shall be understood: the interoperability, and the authenticity of the provenance trails.

Interoperability is an important aspect, as a lesson from e-Science, because it is (both practically and theoretically) impossible to require all WMSs to use the same provenance format. We can reuse the mechanism followed by e-Science research, to use PROV-O [16] as the base level interoperability mechanism across different provenance schemas. In addition to that, dedicated provenance ontology as a common base may be developed specifically for the nature of Web Science research.

Authenticity is necessary to keep the record away from being intentionally/maliciously tampered with, due to the multiplicity of such services. In a context where reputation matters, this is less of an issue, such as when the services are provided by research institutions. However, in other contexts, appropriate cryptographic approaches shall be employed, such as digital signatures or even blockchains. Essentially, every record should be verifiable, guaranteeing the retrieved version is the version being initially stored.

In addition, if the data consumer / researcher is questionable, authenticity mechanism should be extended to ensure the provenance being sent to the storage is the same provenance being collected during workflow execution. If the provenance storage collates with the workflow engine, this can be guaranteed automatically by directly transmitting provenance from the workflow engine to the provenance store. If otherwise, the provenance storage may require the workflow engine to be executed inside a secure environment as a means to minimise issues from user inputs (Sec 4.2).

We do not hide that, eventually, some level of trust in the data consumers is needed, as we aim to provide an *auditing layer* rather than an enforcement framework. Misbehaviours should be **discoverable** from the auditing layer, rather than prevented in advance, as is impossible nowadays with humans monitoring everything.

The situation for the experiment registry is similar to provenance storage, but without executions from the data consumer side. We do not discuss that separately.

4.2 Confidential processing

When necessary, stronger confidentiality or integrity guarantee can be provided by confidential computational mechanisms. If a single source of trust can be established, running computation in a Trusted Execution Environment (TEE) [12] is a sensible option that not only guarantees confidentiality, but also genuineness of the processing through remote attestation [2]. However, in a diversified

trust environment, TEE is not appropriate, while (secure) multi-party computation (MPC) [8] is a promising technology for such contexts. Specifically, Libertas [23] provides a flexible framework for running MPC with a large number of data providers across the Web, as a promising solution for the Web Science context.

4.3 Policy engine and policy language

As illustrated above, the policy engine needs to perform two types of reasoning, which is not a common feature for all policy languages. Previous literature suggested the distinction between policy languages, to be “perennial” or non-perennial [24], where two critical factors are the ability to perform policy derivation and checking policies for downstream tasks. Therefore, research along the line of “perennial policy” should be most appropriate for this framework, such as psDToU [24], Dr.Aid [11], Smart object [13], and CamFlow [10]. Among them, psDToU may be the most promising language for its rich expressiveness among all such languages, also taking into consideration different (but alignable) vocabularies. Further extensions of the language are expected to include further types of information, as well as evaluating its appropriateness in the context of Web Science.

5 Adoption path and open questions

Apart from the utility and technical viability, its adoption is also an important question for any new technology.

In Sec 3, we have discussed how the proposed framework may benefit all stakeholders involved in the research cycle, and their roles in the framework, covering the data providers, the researchers / data consumers, research institutions, and research registries such as OSF; we also discussed how relevant research practices may be reformed and supported in the framework, such as ethics clearance and verification.

This not only has visionary benefits, but also short-term to mid-term practical benefits, even if partially adopting the framework. We have discussed the benefits for research institutions and researchers in the topic of ethics above, covering application, approval, communication, and (posterior) demonstration/verification. The reduction of (human) resource requirements and reputational benefits can form incentives for research institutions to adopt this alternative paradigm, to support provenance storage, workflow-based experiment descriptions, and formal policies (and automated policy reasoning). They may also want to provide institution-wide workflow engines (e.g. as an addition to their computing clusters) for the sake of providing better service, easier collection of provenance, and assurance in the compliance during experiment execution.

Experiment registries can benefit from the structured formal details of experiments as scientific workflows, and support advanced analysis, such as drawing comparisons and warning of potential problems. This can improve and diversify the services they provide to researchers, and incentivise the adoption by the registries.

Benefits of data sources have also been discussed in Sec 3, for better accountability and improved trust in researchers. This reduces the burden of the access application system, and the compliance checking system, while also streamlining with the logging system –

data access application and compliance checking are conducted simultaneously and supported by the (semi-)automated policy reasoning; access logs and experimental information are stored together to facilitate ex post compliance analysis.

Of course, having these incentives alone may not constitute the practice in adoption, due to different socio-technical questions. Apart from technical viability and incentives, the adoption burden itself may be worth further investigation. That includes societal burdens, such as the inertia for keeping existing practices, which cannot be answered by technology, or skill burdens, such as the understanding and technical support for researchers in different domains to use this framework. Establishing and managing trust between different actors is also an open question, where partial solutions may rely on technologies discussed above. For the technical side, ideally, a common protocol (or a set of interoperable protocols) is agreed upon for the framework as a basis, which would require further development and negotiation. A consensus or interoperability mechanism across different workflow modelling and provenance representation should also be established.

Finally, this framework works on the assumption that FAIR is achievable for data, experiments and logs. But we are aware there are practical burdens for many Web Science research to be fully FAIR, such as social media analysis, where scraping may be performed. This has reasons in both the technical challenge of social media platforms not (or closing down [7]) data access endpoints, and the potential governance and policy implications for providing archives by the researchers. We anticipate that this framework shows a better protection and compliance of data and associated policies, lowering the concerns of non-compliance. This hopefully could incentivise users who generate these contents, and further accelerate regulatory requirements for programmatic research access to such data sources, similar to EU Digital Services Act [3].

6 Summary

This paper proposes the design of a technical framework for auditable Web Science research, especially subfields requiring extensive data harvesting, focusing on the reusability and data-usage policy compliance, taking inspiration from other scientific fields, namely e-Science and data usage control. The proposed framework has benefits for all actors involved in such research. We further analyzed its other features, as well as its applicability to recent exploration of LLM research and agents; some technical possibility of achieving the framework is also examined.

In general, we anticipate this framework, as a viable technical solution, to foster discussions on ethical while low-burden data collection and processing across the Web or other decentralized contexts, which is often an overlooked topic in Web Science or research in general. This should support long-term sustainable Web development and research, and create an example for practical and ethical research or data processing for other domains to follow.

References

- [1] Malcolm Atkinson, Sandra Gesing, Johan Montagnat, and Ian Taylor. 2017. Scientific Workflows: Past, Present and Future. *Future Generation Computer Systems* 75 (Oct. 2017), 216–227. doi:10.1016/j.future.2017.05.041
- [2] Manuel Barbosa, Bernardo Portela, Guillaume Scerri, and Bogdan Warinschi. 2016. Foundations of Hardware-Based Attested Computation and Application to SGX. In *2016 IEEE European Symposium on Security and Privacy (EuroSP)*. IEEE, Saarbrücken, 245–260. doi:10.1109/EuroSP.2016.28
- [3] European Parliament and Council. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act). <https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng>.
- [4] Melanie Feinberg, Will Sutherland, Sarah Beth Nelson, Mohammad Hossein Jarrahi, and Arcot Rajasekar. 2020. The New Reality of Reproducibility: The Role of Data Work in Scientific Research. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 035:1–035:22. doi:10.1145/3392840
- [5] Anthony J. G. Hey (Ed.). 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington.
- [6] Luis-Daniel Ibáñez, John Domingue, Sabrina Kirrane, Oshani Seneviratne, Aisling Third, and Maria-Esther Vidal. 2023. Trust, Accountability, and Autonomy in Knowledge Graph-Based AI for Self-Determination. *Transactions on Graph Data and Knowledge* 1, 1 (Dec. 2023).
- [7] Kai Kupferschmidt. 2023. Twitter's Threat to Curtail Free Data Access Angers Scientists. *Science* 379, 6633 (Feb. 2023), 624–625. doi:10.1126/science.adh1696
- [8] Yehuda Lindell. 2020. Secure Multiparty Computation. *Commun. ACM* 64, 1 (Dec. 2020), 86–96. doi:10.1145/3387108
- [9] Spencer C. Norris, John S. Erickson, and Deborah L. McGuinness. 2017. A Semantic Workflow Approach to Web Science Analytics. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*. Association for Computing Machinery, New York, NY, USA, 403–404. doi:10.1145/3091478.3098875
- [10] Thomas F. J.-M. Pasquier, Jatinder Singh, David Eysers, and Jean Bacon. 2017. CamFlow: Managed Data-sharing for Cloud Services. *IEEE Transactions on Cloud Computing* 5, 3 (July 2017), 472–484. doi:10.1109/TCC.2015.2489211 arXiv:1506.04391
- [11] Rui Zhao, Malcolm Atkinson, Petros Papapanagiotou, Federica Magnoni, and Jacques Fleuriot. 2021. Dr.Aid: Supporting Data-governance Rule Compliance for Decentralized Collaboration in an Automated Way. In *The 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*. doi:10.1145/3479604
- [12] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. 2015. Trusted Execution Environment: What It Is, and What It Is Not. In *2015 IEEE Trustcom/BigDataSE/ISPA*, Vol. 1. 57–64. doi:10.1109/Trustcom.2015.357
- [13] Gokhan Sagirlar, Barbara Carminati, and Elena Ferrari. 2018. Decentralizing Privacy Enforcement for Internet of Things Smart Objects. *Computer Networks* 143 (Oct. 2018), 112–125. doi:10.1016/j.comnet.2018.07.019
- [14] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated Delegation and Authorized AI Agents. *CoRR abs/2501.09674* (2025).
- [15] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359, 6380 (March 2018), 1146–1151. doi:10.1126/science.aap9559
- [16] W3C. 2013. PROV-O: The PROV Ontology. <https://www.w3.org/TR/2013/REC-prov-o-20130430/>.
- [17] W3C OWL Working Group. 2012. OWL 2 Web Ontology Language Document Overview (Second Edition). <https://www.w3.org/TR/owl2-overview/>.
- [18] Daniel J Weitzner, Jim Hendler, Tim Berners-Lee, and Dan Connolly. 2006. Creating a policy-aware web: Discretionary, rule-based access for the world wide web. In *Web and information security*. IGI Global, 1–31.
- [19] Mark D. Wilkinson, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (March 2016), 160018. doi:10.1038/sdata.2016.18
- [20] Sean R Wilkinson, et al. 2025. Applying the FAIR Principles to Computational Workflows. *Scientific Data* 12, 1 (Feb. 2025), 328. doi:10.1038/s41597-025-04451-9
- [21] Jesse Wright. 2025. Data Wallets. <https://blog.jeswr.org/2025/02/14/data-wallets>
- [22] Jesse Wright, Beatriz Esteves, and Rui Zhao. 2024. Me Want Cookie! Towards Automated and Transparent Data Governance on the Web. In *NXDG, 20th International Conference on Semantic Systems (SEMANTiCS)*. Amsterdam. doi:10.48550/arXiv.2408.09071 arXiv:2408.09071 [cs]
- [23] Rui Zhao, Naman Goel, Nitin Agrawal, Jun Zhao, Jake Stein, Ruben Verborgh, Reuben Binns, Tim Berners-Lee, and Nigel Shadbolt. 2025. Libertas: Privacy-Preserving Collective Computation for Decentralised Personal Data Stores. doi:10.48550/arXiv.2309.16365
- [24] Rui Zhao and Jun Zhao. 2024. Perennial Semantic Data Terms of Use for Decentralized Web. In *Proceedings of The ACM Web Conference 2024*. ACM, Singapore, 2238–2249. doi:10.1145/3589334.3645631