From policy to practice in data governance and responsible data stewardship: system design for data intermediaries

Powar, Jovan; Janssen, H.; Cloethe, Richard; Singh, Jatinder

DOI
10.1145/3715275.3732164

Publication date
2025

Document Version
Final published version

Published in
ACM FAccT '25

From policy to practice in data governance and responsible data stewardship: system design for data intermediaries

Jovan Powar
University of Cambridge
Department of Computer Science and Technology
Cambridge, United Kingdom
Alan Turing Institute
London, United Kingdom
jovan.powar@cst.cam.ac.uk

Heleen Janssen
University of Amsterdam
Institute for Information Law
Amsterdam, Netherlands
h.l.janssen@uva.nl

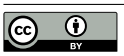
Richard Cloete
Harvard University
Center for Astrophysics
Cambridge, USA
richard.cloete@cfa.harvard.edu

Jatinder Singh
University of Cambridge
Department of Computer Science and Technology
Cambridge, United Kingdom
University of Duisburg-Essen
Research Centre Trust, UA Ruhr
Duisburg, Germany
jatinder.singh@cl.cam.ac.uk

Abstract

Data intermediaries are a family of approaches to data governance that aim to rebalance power between third party data users and those to whom data relates or 'belongs'. These approaches—such as data trusts, commons, institutions, and personal data stores—construct a governance regime around the management and use of data by introducing an 'intermediary' party. Data intermediaries are gaining significant traction in the socio-legal data governance community, and attracting increasing attention from regulators. However, so far the study of data intermediaries predominantly focuses on their policy, organisational, and economic aspects, while the technical design and implementation of systems that operationalise intermediary data governance have gone under-addressed.

In this paper, we initiate the technical study of data intermediaries as a system design problem. We first review the existing literature on intermediaries to catalogue the various proposed forms of intermediaries and their purposes, extracting a set of design questions which produce a specification of the required intermediary system. We then present a design language of system architectures, which describe the spectrum of practical forms that intermediary ecosystems can take. Using this language we illustrate the challenges of designing intermediary systems: we show how governance goals can be achieved or supported by architecture choice; we show how architectures support different governance approaches; and we discuss the drawbacks and challenges that arise from certain design choices. We further provide a technical agenda for data intermediaries, highlighting implementation methods, relevant trends, and opportunities for cross-disciplinary collaboration.



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732164

CCS Concepts

• **Information systems** → **Data management systems**; • **General and reference** → **Design**; *Surveys and overviews*; • **Computer systems organization** → *Distributed architectures*; • **Security and privacy** → *Information accountability and usage control*; Trust frameworks; Human and societal aspects of security and privacy.

Keywords

data governance, data trusts, data commons, data cooperatives, data institutions, responsible data management, Personal Data Stores, socio-technical systems, structured design, distributed architectures

ACM Reference Format:
Jovan Powar, Heleen Janssen, Richard Cloete, and Jatinder Singh. 2025. From policy to practice in data governance and responsible data stewardship: system design for data intermediaries. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715275.3732164

1 Introduction

Across disciplines, new approaches [32, 61, 81] for managing and governing data use are emerging in response to increased concerns around the scale and opacity of data collection, processing, and sharing [60, 81, 115, 120]. These approaches aim to rebalance the interests and power dynamics of data ecosystems in favour of those to whom data relates or 'belongs', who previously had limited involvement in the use of that data, a lack of control that is reinforced by systemic power and information asymmetries [22, 23, 68].

These power dynamics are becoming increasingly salient and difficult to manage in the era of AI, particularly with foundation models and 'Generative AI', which require vast training data which is often obtained with dubious legality and minimal transparency [28, 116]. This has profound costs for individuals, private entities, and the digital commons at large [63]. As the demand for

quality data for AI models increases, it is imperative we provide new and effective measures to govern the data supplier-recipient relationship [67].

While approaches vary, a prominent theme is the idea of intermediated data processing. This is the idea of introducing an entity, the **data intermediary** [66], that is charged with creating and effecting a management and governance regime around the data of **data suppliers**, and its processing in the interests of **data recipients**. This governance regime guides, constrains, and monitors data use to ensure it remains in line with the suppliers' rights and interests.

Intermediaries are gaining attention and traction in practical trials [4, 53] as well as among governments [51, 83] and legislators [96]. However, to date most work on intermediaries has occurred within legal, policy, and social/public sector research communities, *without direct consideration of their technical aspects*. This is a critical omission: intermediaries are not simply socio-legal constructions implemented by technical systems, but socio-technical systems in which practical system design is the fundamental basis for any governance regime implemented using technical, legal, and/or other mechanisms. If intermediaries are to become a successful model of data governance, we require methodologies and guidance around their system design, and an understanding of the connection and interaction between that design and governance outcomes.

With this paper we provide a concrete foundation for the system design of data intermediaries, and begin to bridge the gap between technical data management and socio-legal data governance communities in both research and practice. Our contributions are:

(1) to extract a set of design questions for specifying intermediaries, through a systematic survey and analysis of the existing intermediaries landscape;
(2) a high-level design language for intermediary ecosystems;
(3) a minimal API that practitioners can use to implement any role within an intermediary ecosystem, and thereby quickly prototype different architectures;
(4) a mapping of intermediary design approaches to their design specification, and a discussion of the complexities and depth of that system design problem;
(5) an identification of bodies of technical work which can be starting points for adapting approaches to particular intermediary architectures, and an agenda for technical research going forward.

2 Background and scope

Across the field of data governance, models of data management that involve an intermediary have arisen in a number of distinct forms. These are a response to concerns around the collection of data [60, 81, 115] about or in some way belonging to individuals or organisations (suppliers), by organisations which harvest value and benefit from that data unto themselves [7, 68, 139]. The expectation that the use of suppliers' data will align with their rights and interests is low, due to the opacity of dataflows, concerns around datafied surveillance, and the systemic power and information asymmetries inherent in current data processing ecosystems [22, 23, 32, 81, 101].

The data governance literature offers a diverse range of systems proposed to effect governance over the use of data by third parties. Some of these proposals have generated rich bodies of work and practical application, while others are largely theoretical or in a nascent phase. Even among models of governance which introduce an intermediary party, we have a range of competing and overlapping terminologies, and numerous efforts at systematising their concepts. This paper does not attempt yet another systematisation of these terms—we review existing mapping exercises in §2.3—instead we adopt definitions from existing work and provide a perspective on system design that cuts across the spectrum of terms and concepts. In this section we situate our work within this literature: we introduce the range of intermediaries already proposed, state the definitions we adopt, and contextualise our work against existing mapping and conceptualisation efforts.

2.1 A brief introduction to data intermediaries

Before we delve into the conceptual details of data intermediaries and related models of data governance, we will provide a brief introduction to the trends and drivers of this nascent field.

Examples of data governance approaches that involve an intermediary party can be found in scholarship across multiple disciplines, as well as under a number of different terminologies, some of which have attracted their own distinct bodies of work [26, 32, 48, 59, 60]. In practice, a number of data governance projects have implemented systems to operationalise these governance approaches [27, 42, 53, 83]. Moreover, intermediaries are attracting attention from legislators and regulators, such as in the EU's Data Governance Act [96].

2.1.1 Use-cases of data intermediaries. Intermediaries have been considered and trialled across a wide range of data ecosystems, tasks, and stakeholder relationships. Data suppliers may be individuals or organisations, public or private, and so the governance goals and design of intermediaries vary widely.

Examples of data intermediaries can be found in scholarship across multiple disciplines and under a number of different terminologies, along with distinct bodies of work [26, 32, 48, 59, 60]. Similarly, a number of data governance projects have implemented systems that can be understood as data intermediaries [27, 42, 53, 83]. We next explore *key classes of data intermediary*.

Some intermediaries already in operation aim to **govern existing relationships or uses of data.** These intermediaries, often proposed or implemented as data trusts, aim to counter the consolidation of power by corporate data and data-driven business models, by providing more agency to suppliers (individuals or organisations) [32]. An example is the case of Worker Info Exchange, which extends the model used by its founders in a lawsuit against Uber, resulting in a win at the UK Supreme Court in favour of drivers—their approach is to facilitate subject access requests by workers, and pool that data in order to act in their collective interests, under the framework of a data trust [40, 52]. Another example is the proposal for a novel governance regime around language data in the context of Large Language Models (LLMs) [67]. This approach involves distributed repositories of data which fall under the umbrella of a 'Data Stewardship Organisation' which co-ordinates a range of organisational frameworks for aligning data management and use with a set of commonly agreed values, including as a counter to power imbalances.

Another class of intermediaries aims to leverage suppliers' data to **enable or improve approaches to known problems**. These may be entirely community-driven, with community members pooling their data to advance shared interests or the broader common good [61]. Other initiatives involve the public sector. This may involve citizen or consumer data being contributed to a public sector body to support policy-making processes and provision of public services [83]. In the other direction, governmental bodies also employ intermediaries to make public data available to enable business decisions and innovation: for example, in the UK the Environment Agency manages a data integration project [2] that collects open data from local authorities and provides data on flooding risk to public and private recipients via online tools, such as for long-term planning for businesses and construction.

A number of intermediary initiatives allow individuals to **support scientific research**. In the digital space, the Open Source Data Donation Framework provided a tool for researchers to recruit and manage individuals contributing their 'digital trace' [4], while the Device Analyzer project [127, 128] offered a single Android app for collecting traces which would be provided under license to academic or public-interest researchers. In health, the Open Humans intermediary enables individuals and communities, such as those with a particular health condition, to donate personal data for use by researchers or other health initiatives [39, 42].

There is also a broad class of intermediaries enabling **entirely new uses of data** for private *and* public interests. Some intermediaries may aim to provide direct benefits to data suppliers by enabling them to monetise or otherwise extract value from their data [87]—this is one of the cases covered by the EU Data Governance Act [96]. One proposed novel use of intermediaries [138] describes a use of privately-held personal data for public tasks: to support mobility safety after a traffic accident in the vicinity of a large football stadium, dynamic parking lot data might provide the responsible public authorities with accurate information helping them to divert traffic away from the incident. Using an intermediary would help to prevent personal data from being transferred to the authorities in the interests of the subjects' trust and privacy, instead safely requesting and processing relevant data from the parking lot owner, and transfer accurate information to the variety of public authorities involved.

Data intermediaries have also been proposed to enable the use of commercially sensitive data across suppliers who are competitors and which otherwise would not share any data. For example, one proposal [123] would serve airlines as suppliers, and manufacturers or service contractors, regulators, and perhaps also airlines as recipients: in order to predict material ageing and damage to aircraft fleets, the sensor data across hundreds of airlines could be consolidated under the governance of an intermediary. This would provide actionable insights without airlines directly disclosing commercially sensitive information to competitors.

2.1.2 A technical operationalisation gap. The examples above show the wide utility of intermediaries across use-cases and datatypes, and the substantial interest across both public and private stakeholders. However, the existing landscape of uptake comes largely from legal and organisational trials, with relatively simple technical means of exerting control. While some of the examples we discuss above originate in technical fields, there remains a gap in technical expertise or guidance around operationalisation.

In a survey we performed in 2022, only two projects came from strictly technical backgrounds: the OSD2F [4], which provided a simple data collection tool for researchers, and Device Analyzer. Each technical project focused on the collection aspect of the intermediary, and did not provide technical means of governance, and both have since been shut down. While the OSD2F was only a collection tool, the Device Analyzer project [127, 128], which was maintained in part by one of this paper's authors, operated as a complete intermediary, providing access to third party recipients. Even then, the technical detail of the project was limited to the data collection process, with the governance regime operationalised solely by legal and organisational means, relying on institutional legal processes and compliance burdens to control processing. Significant operational frictions introduced by the hands-on data transfer process and institutional contractual relationships limited the scalability of the sharing where sensitive data was involved. Updating the infrastructure to leverage greater automatability and technical governance tools was considered but at the time there was a lack of existing technical practice to serve as a reference.

Another of the examples above, Open Humans [39, 42] is a good example of a technically mature infrastructure, providing a portal for suppliers and recipients and a notebook-style platform for data analysis. While the platform provides some basic mediation of the supplier-recipient relationship through dataset- and project-level access permissions, oversight of the actual work done with data is limited to manual oversight or legal compliance.

These projects demonstrate that while models of distributed data collection are a reasonably mature field of practice, the governance of the data analysis process that occurs thereafter—the interface between the data collector and the prospective recipient, and the constraint of the recipient's usage so that it accords with the data supplier's rights and interests—lacks a technical body of work. This is a particular barrier to projects where suppliers or their data (e.g. digital traces) are particularly sensitive to breaches of the governance regime. In these cases, the project is forced to rely wholly on socio-legal governance tools such as contracts, even where it is clear that incorporating some technical methods to mediate or instrument data use may allow for a similar level of protection at relatively lower operational costs.

2.1.3 Legal and regulatory drivers. The development of new models of data governance has also been driven by interest from legislators and regulators [25]. In fact, the recently adopted EU Data Governance Act (EU DGA) [96] explicitly purports to account for data intermediaries, by introducing particular articulations of intermediated data processing and its governance. In terms of technological sovereignty, it also represents a strategic bet on the future direction of the data economy, by proactively embracing key enabling technologies and infrastructures [25]. Tridgell et al. [122] discuss the extent to which 'stewardship' is finding uptake in another area of legislation—the EU's 2024 Cyber Resilience Act legislates for intermediary-like entities to govern open-source software. This momentum underscores the urgency of integrating the technical community into the study of data governance, not only to contribute implementation detail, but to support regulation and policymaking.

2.2 Definitions

Data intermediaries are entities introduced to govern the use of data within an *ecosystem* comprised of multiple stakeholders, each of which has an interest in or plays a role in the data's use [66, 81]. The object of the governance regime is the action of this ecosystem as a whole—for the remainder of this paper we will refer to **instances of intermediated processing**, undertaken by a **data intermediary ecosystem**, of which the data intermediary is one party. In this paper we will follow the definitions given by Janssen and Singh [66] who identify three key types of stakeholder that make up a data intermediary ecosystem[1]:

data suppliers those individuals, communities, or enterprises that make their data available for processing, who have some rights to that data, and in whose interests the governance regime is to act;

recipients those interested in using or otherwise leveraging supplier data, i.e. those who would be the direct beneficiary or instigator of the instance of intermediated processing; and the

data intermediary which serves as a mediator between the suppliers and each recipient [66]. The intermediary works to govern the data in specific ways, and provides the suppliers some degree of confidence that the use of their data will be according to their rights and interests.

It is crucial to note that models of data governance that fit this definition do not effect governance simply through the construction and operation of the intermediary. Instead, the governance regime is achieved through the operation of a **socio-technical system across the parties in the ecosystem**. Creating this governance regime involves introducing the intermediary as an additional party to the processing task, but also assigning roles and actions to the suppliers and recipients. In practice, this system is operationalised through technical infrastructure, contractual and legal obligations, and operating specifications such as data formats, methods of communication, and security protocols.

2.3 The landscape of data governance models

Many data governance (or *data stewardship*) models fitting the description of data intermediaries have been proposed, each with a distinct approach to data governance and exhibiting design features, goals, and strategies particular to its context and purpose. These models have been proposed under a number of terminologies such as *data trusts* and *data cooperatives*. Efforts to map these models and align their terminologies already exist: Janssen and Singh [66] and Micheli et al. [80] both systematise approaches under the 'intermediary' terminology; Petreski and Cheong [98] focus on data cooperatives, but explicitly situate that concept within the larger context of governance models (including non-intermediary approaches); Wong et al. [132] do similarly for data commons; and the work of Hardinges and Keller [57, 58] collects an even broader range of governance (stewardship, in their terms) models under the umbrella concept of a *data institution*.

[1]There are, of course, other parties that can be involved in the process exchanging data and supporting computation (including cloud providers or platform infrastructure); however, for this paper we focus on the key actors of the data intermediary ecosystem, around which intermediary governance concerns are targeted.

These mapping exercises do good work to systematise the commonalities and differences across high-level approaches, largely in terms of the high-level decisions or issues they tackle, their approaches to decision-making, and the power dynamics they impose. For example, Hardinges and Keller [57] distinguish governance approaches according to their decision-making: individual, collective, or delegated. An alternative distinction is in the level of participation of data suppliers in governance, termed 'top-down' when the level is low and 'bottom-up' [32, 57, 81] when high.

2.3.1 Other relevant work. The term data intermediaries is also used in the realm of public or 'open' data. Shaharudin and Janssen [109, 110] have published work on these 'open data intermediaries', organising their purposes and the governance activities they perform. While this literature has some crossover with the 'data stewardship' branch of data governance we have discussed above, its focus on open data means the systems considered do not overlap with those we are concerned with—the definition we follow might be considered models of managing access to 'closed' data.

A significant area of related work is in the EU's initiative towards common European 'data spaces' [24, 41]. These can be thought of as high-level organisational structures intended to ease dataflows within a certain sector. Data spaces take a industry-level and market-based perspective on managing the use of data, which is distinct from intermediaries' focus on mediating the relationship between supplier and recipient. As such the systems that operationalise data spaces are more like closed ecosystems comprising a wider range of stakeholders and roles [114]. Atzori et al. [5] provide a thorough introduction to data spaces' concepts and technical requirements and relate them to the data management literature.

2.3.2 Why intermediaries: governance through structure. It should be clear by the number of mapping exercises disambiguating between overlapping terminologies, showing how category x can in fact be framed as a subcategory of y, that choosing to frame a governance approach as a data intermediary is one of many options. We argue that the framing is a useful one, though, since intermediaries present an operational means to implement governance approaches across the spectra defined in those works: intermediaries may be top-down or bottom-up; and may enable individual, collective, or delegated decision-making. Intermediaries represent structural approaches to data governance, whose governance effect is determined by the structuring of the intermediary ecosystem and the practical system that enforces and operates on that structure.

However, if intermediaries are to actually prove a useful model of data governance in practice, we must identify and present the connections between high-level governance goals or concerns to practical, structural system design.

3 The design specification of an intermediary system

The first step in addressing the system design problem of data intermediaries is to identify the decision space that system designers must navigate. In §2.2 we stated that the creation of a governance regime begins by assigning behaviour, functionality, roles, and responsibilities to all actors within the intermediary ecosystem. In

Table 1: Models of intermediaries extracted by analysis (A)

Approach	Key features
Data trust [26, 32, 48, 49, 56, 90, 104]	Takes on responsibility to steward supplier data for agreed purposes; may be based on fiduciary duty to suppliers, and/or a contractual or statutory legal obligations.
Data commons [96, 132]	Members voluntarily 'pool' their data for the benefit of a specific community or the general public.
Data cooperative [61, 96]	An intermediary owned and democratically controlled by its members who delegate control over their data.
Data collaborative [118]	Participants from different sectors, e.g. private companies, research institutions, government agencies, exchanging data and data expertise to help solve public sector problems.
Personal data store [27, 64, 65, 73, 96]	Systems offering suppliers means to mediate, monitor and control how their data is accessed, used, or shared.
Data marketplace [87]	Also termed 'data brokers' or 'trusted third parties' that work to allow the trading of data.

system design terms, we can consider this the *functional specification* of the data intermediary system.

In this section we discuss three analyses we performed on the intermediaries literature to capture: **(A)** the distinct models of intermediaries proposed, **(B)** those intermediaries' purposes, and **(C)** a set of design questions which determine the functional specification. §3.1 discusses our methodology and the results of analyses **(A)** and **(B)**, while §3.2 presents the results of analysis **(C)**.

3.1 Our survey

3.1.1 Methodology. The three analyses we performed followed the stages of thematic analysis set out by Braun and Clarke [17], with one variation: as our work is additive to the existing landscaping works noted in §2.3, we skipped the first three Braun-Clarke stages during analysis **(A)** by prepopulating the codes with the categories already described by Janssen and Singh [66].

We began with a literature base drawn from those two surveys, and grew it by a two-step iterative process: adding literature by database searches of major repositories (e.g. Scopus, Google Scholar, ACM DL) and snowball sampling, and updating analyses **(A)** and **(B)**; and repeating until the literature and themes were stable. The final set of search terms we arrived at was 'data {sharing | protection | steward | commons | trust | cooperation | collaborative | ownership | broker | institution }', 'fiduciary duty', 'empowerment' and ' {personal | user} {privacy | data} management'.

*3.1.2 Results of analyses **(A)** and **(B)**.* Analysis **(A)** was intended to ensure that the literature base fully covered the range of models under the intermediaries umbrella, which was achieved (Table 1). Analysis **(B)** aimed to identify the breadth of purposes for which intermediaries have been proposed. These, shown in Table 2, are naturally not exhaustive of potential intermediary aims since existing work covers only a fraction of intermediaries' potential use-cases. Over time a variety of other purposes will likely emerge, but the existing breadth demonstrates a notable degree of uptake already, as well as convergence around particular aims.

3.2 Governance specification questions

Analysis **(C)** aimed to extract a set of **governance design questions** that cover the range of decisions that must be made when producing the system's specification. The results, shown in Table 3, can be grouped into three categories:

Table 2: Results of analysis (B): known intermediary purposes.

Examples	Intermediary purpose
[26, 32, 48, 49, 96, 104]	Protecting the interests and rights of data suppliers
[48, 49]	Correcting power imbalances by empowering and including data suppliers in setting the terms of data use
[26, 32, 39, 48]	Supporting suppliers in managing their data, e.g managing consent, exercising data rights
[32, 60]	Enabling suppliers to wield collective bargaining power
[9, 87]	Enabling supplier monetisation/value extraction from data
[6, 26, 77, 96, 107]	Allowing the pooling of data for particular aims, e.g. research, investigative journalism, public interest.
[43, 130]	Enforcing compliance with individual rights, e.g. employment rights, anti-discrimination in finance.

Table 3: Results of analysis (C): design specification questions

Category		Governance question	
Task details	***T1***	What purposes should/can the intermediary pursue?	[18]
	T2	What types of data will be involved?	[48, 104]
	T3	Who are the suppliers and recipients?	[15, 49, 59, 104]
	T4	What uses of data are legitimate?	[15, 18, 48, 49, 59, 90]
Structural specification	***S1***	Where and how will data be captured and stored?	[15, 59]
	S2	Who will perform computation on the data?	[18, 49]
	S3	Who transfers data or results to the recipient?	[18, 49, 59]
	S4	How will tasks and responsibilities be distributed between parties in the ecosystem?	[15, 49, 92]
Governance strategies	***G1***	How will oversight and accountability be effected?	[92, 132]
	G2	How will rules and boundaries be actively enforced?	[49, 92]
	G3	How will risks of conflict and divergent incentives between parties be identified and managed?	[90, 92, 132]
	G4	How will governance decisions be accounted for?	[18, 48, 59]

Task details capturing the basic details of the processing task that is to be governed: the data and parties involved, the purpose of processing, and the extent of 'legitimate' use.

Structural specification fleshing out the tasks and responsibilities assigned to each role in the ecosystem.

Governance strategies specific strategies for how governance requirements will be achieved, through policy, compliance, or the application of technical or procedural tools.

For a practitioner given the task of designing an intermediary ecosystem, some of these questions can be answered in part or wholly by reflecting on the context at hand, while some may require further interpretation or a decision-making process. For example, in a case where an intermediary is proposed to manage health data, the question of data type (***T2***) is already prescribed, and in many jurisdictions the question of how data can be legitimately used (***T4***) will already be partly answered by regulations (such as HIPAA in the USA). However, questions such as who will perform computation (***S2***) and where data is stored (***S1***) may be open to deliberation—different approaches may be considered and design decisions across the space of governance questions traded off against each other.

3.3 Implementing the specification

In the literature, the governance specification questions are largely discussed in terms of high-level goals and principles—discussing why a certain approach (i.e. a particular answer to one of the questions) would be beneficial to certain parties. However, the prospective practitioner still lacks a cohesive approach to answering all of the questions together, and exploring the potential socio-technical systems that would create and enact the desired governance regime.

It is important to note that the socio-technical system of an intermediary ecosystem is dynamic—an answer to one of the questions ***S1–4*** or ***G1–4*** has implications for *all of the others*, constraining the options that a practitioner has to achieve those aspects of the specification, or indeed supporting and creating new options.

3.3.1 Mechanisms of governance. Governance is performed by the intermediary and the ecosystem through a range of mechanisms—legal and normative as well as technical and organisational. Most discussions to date focus on the legal: typical mechanisms include contractual mechanisms that legally bind the contracting parties to do or to tolerate something, such as a supplier permitting the intermediary to store that supplier's data in the intermediary's database, or where the intermediary performs computation over the supplier's data, according to bespoke contractual terms [26, 81, 104].

In addition to contracts, fiduciary duties of undivided loyalty and care might apply [26, 32, 37, 96], requiring that the intermediary in every situation acts in the best interests of the suppliers as best judged for their known preferences and the context. Fiduciary duties are sometimes considered an extra safeguard to the intermediary's contractual duties, and aim at creating predictable data supply chains with increased accountability towards the suppliers [49]; other authors strongly advocate for establishing the fiduciary duty as an essential element of the intermediary's relationship with data suppliers [32]. Another mechanism enabling a data intermediary's governance of supplier data under agreed terms is consent, where personal data is involved (see §4.4.3).

Technical and operational mechanisms for data governance have been less-considered as part of this literature, though it is clear that there are a range of mechanisms that would be applicable to intermediaries, such as secure multiparty computation [76], data provenance [113], and privacy-enhancing technologies [102, 103]. Such mechanisms not only dictate the capabilities of parties within the ecosystem and provide concrete interventions, but can also shape power dynamics between parties or influence their behaviour—for example through communication rates and protocols [113].

3.3.2 Our approach: structure design. To connect the diverse universe of mechanisms that can be employed to perform data governance to the goals and principles of a specific task, we will rely on our insight from §2.3.2—the uniting principle of data intermediaries as a governance model is that they are fundamentally structural. In the next section we present our system design language for data intermediaries, which is based on the approach of structure design in software engineering.

4 A design language for intermediary ecosystems

An intermediated data processing task is performed as a collective operation between the data suppliers, the intermediary, and the recipient. As the foundation for our approach to intermediaries as a systems challenge, we introduce a design language of intermediary **architectures** to express the range of practical arrangements by which this can occur, derived from our literature survey.

4.1 Methodology

The primary design feature of data intermediaries as an approach to data governance is the distributed structure imposed onto the processing task. To capture this, we adopt a structured design approach as used in software architecture engineering.

Structured design involves decomposing a system into modules, defining each module's responsibilities and interfaces, and specifying the connections between them which govern their interactions [136]. Adopting this approach, we are able to identify each component in a task of intermediated data processing and isolate it as a potential surface for applying governance mechanisms. The decomposition of a system into modules is led by three main principles [46, 111]: separation of concerns, which dictates that each module addresses a particular concern (here being governance decision-points and inter-party communications), low coupling (meaning low interdependence between modules), and high cohesion (all elements of a module belong together).

Naturally, the first layer of decomposition we performed was by actor. Then, applying the above principles, we further broke down and specified each actor's behaviour until we arrived at a language of distinct architectures composed of orderings of activities performed during the intermediated processing task.

4.2 Activities

The first part of our language is the **activities** that parties perform during intermediated processing, which fall into three categories:

Q – Request One party asks another to perform some work.
P – Processing A party processes data it holds.
R – Release One party communicates some data to another.

We also define an activity—**DCS (Data Capture and Storage)**—that occurs outside the lifecycle of any one processing task, as a catch-all for any 'setup' activities that are performed as prerequisites to the ecosystem's operation. *All* intermediated processing tasks can be expressed as a sequence of these activities.

A **request** may vary in content from a description of work in natural language with the expectation that it will be read by a human, to a machine readable query or executable program, depending on the nature or capabilities of the party on the receiving side,[2] or as an intentional design decision to force manual human oversight.

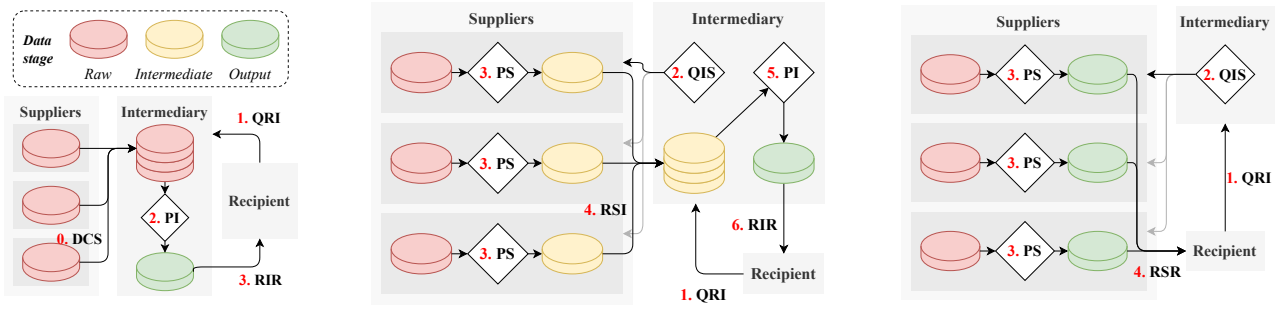
Processing is defined broadly—as in data protection regulation such as the GDPR [97]—to include most anything done to or with data, covering the retrieval of data from storage as well as computation on that data. We do not include the processing performed by the recipient, as this occurs outside the lifecycle of an intermediary.

Release activities describe the flow of supplier information, either in the form of 'raw' data or 'intermediate' data produced by processing raw data, excluding communications like handshakes or metadata exchange. Access to this information is the fundamental concern of the governance regime. We note release separately from processing to capture the specific phenomenon of data leaving a party's direct control, a crucial point for governance intervention.

[2]Recall that we consider intermediary ecosystems as socio-technical systems, and so human participation is considered part of the activity. In fact, we make no assumption that a computerised or digital system is used to perform any activity in an intermediary ecosystem (though in practice it most often will).

Table 4: Activities that occur during intermediated data processing.

Category	*Data storage*	*Request*		*Processing*		*Release*		
Code	***DCS***	***QRI***	***QIS***	***PI***	***PS***	***RSI***	***RIR***	***RSR***
Description	Data is captured and stored, including distribution to intermediary if appropriate	Recipient initiates a task	Intermediary instructs suppliers	Intermediary performs processing	Supplier performs processing	Suppliers release data to intermediary	Intermediary releases data to recipient	Supplier releases data to recipient



(a) Centralised **(b) Mixed** **(c) Conductor**

Figure 1: The three intermediary architectures. Arrows denote communication, and DCS in Mixed and Conductor are omitted for clarity.

4.3 Architectures

Using activities as building blocks, we will now present a generic set of **intermediary architectures**, activity sequences which represent the various ways in which a processing task may be performed by the ecosystem. Each architecture represents a different approach to intermediary design in terms of where data, processing, and control are concentrated. These are not simply technical details, but *reflect and reinforce* the roles and capabilities of the parties, and determine what governance strategies can be employed and to what extent. We describe three major architectures spanning a spectrum of centralisation, shown[3] in Fig. 1:

Centralised One extreme of the continuum, in which data resides and processing occurs entirely at the intermediary during the processing lifecycle and suppliers are not involved contemporaneously to any instance of processing.

Conductor The other extreme, processing is performed solely at the supplier (e.g. via a Personal Data Store [53, 64, 78, 124]) and released directly by the supplier to the recipient. The role of the intermediary is to co-ordinate between parties (e.g. as a trusted central authority in key exchange), and it holds no data.

Mixed In the space between the two, systems in which some processing is performed by both the intermediary and its suppliers.

Mixed architectures combine aspects of the Centralised and Conductor architectures; for example, providing live involvement of suppliers while performing processing tasks that would be difficult or inefficient to distribute across a heterogeneous set of suppliers, or perhaps as a transparency mechanism. As the Mixed architecture spans a range of different centralisation-decentralisation balances, we can further subcategorise it into three sub-architectures:

Data-on-demand As in Centralised, all non-trivial computation occurs at the intermediary (during **PI**), but not all required data is held by the intermediary, though some may be; data must be fetched from suppliers (i.e. **PS** is simply retrieval).

Aggregator As in Conductor, the majority of computation on suppliers' data is performed by the suppliers, but the intermediary collects the results before release to the recipient. There is no direct communication between suppliers and recipient. *Map-reduce* programming models are examples of Aggregator architectures.

True-split Non-trivial computation occurs both at the suppliers and the intermediary. Most Secure Multi-party Computation (SMPC) algorithms can be thought of as True-split architectures: e.g. consensus algorithms which designate a single aggregator for votes [71, 91]; or federated learning algorithms where participants (suppliers) contribute incrementally to a model held by the central party (the intermediary) [70].

This list of architectures is not exhaustive of all multiparty processing models involving a trusted (or partly-trusted) third party, but covers all intermediaries proposed or hypothesised in existing work, providing the best starting position for researchers and practitioners to explore the design considerations of intermediaries.

4.3.1 A minimal API for ecosystem parties. By breaking down intermediated processing into a discrete number of activity categories that are common to all parties, we have produced a minimal API (a set of functions that a technical system exposes to others) that enables an entity to act as any party in the intermediary ecosystem. This allows architectures to be easily prototyped in a modular fashion, and each API function presents a distinct practical opportunity to apply governance interventions. Table 5 shows this minimal

[3]Many models of multiparty computation require some back-and-forth between intermediaries and suppliers, for example for multiple rounds of parameter voting. These would be represented as activity cycles (**QI**–**PS**–**RSI**). Fig. 1 omits these cycles for clarity. Similarly, deliberative processes where a response to the recipient's proposed processing task is formulated collectively by suppliers are omitted.

Table 5: The minimal generic API for an ecosystem entity. A technical system that implements this basic set of functions will be able to play any of the three roles in the intermediary ecosystem.

Function	Purpose
Connect	Connect to a local data store, establish a connection to the intermediary, and begin listening for incoming messages.
Register	Register the party with the data intermediary.
Request	Send a request to a specific party
Receive	Receive messages from other parties.
Parse	If a request is received, determine whether data needs to be retrieved from a database, computation needs to be performed, or whether the request needs to be forwarded to another party. If there is a policy to be applied to incoming requests, *ApplyPolicy*.
DoCompute	Perform a computation on data. Instructions may be results of prior *Parse*. If party has a policy on Processing, *ApplyPolicy*.
ApplyPolicy	Perform any active governance measures, e.g. check recipient or task-level permissions against own or intermediary's database/policy.

API, which we validated by implementing toy simulations of each architecture to conduct data analyses on Kaggle datasets[4].

4.4 Governance considerations of architectures

The architecture is the fundamental design feature of an intermediary ecosystem, determining the feasibility of governance strategies, the roles each party will play, and providing a scaffold onto which tools from technical and non-technical disciplines can be added. We will now develop the design problem posed by intermediaries by exploring how architecture choice affects the governance regime and might constrain or enable further design decisions, with reference to the specification questions in Table 3. These considerations are not exhaustive but represent crucial areas that will require attention in a successful body of technical practice and research.

4.4.1 Degrees of centralisation. Data and computation may be distributed between intermediary and suppliers in numerous ways—this is the focus of governance specification questions ***S1***, ***S2***, and ***S4***. While the merits of decentralisation are well-studied from technical and governance perspectives [3, 65, 85], we note the specific ways in which it affects the governance function of intermediaries.

Locus of data. Deciding where sensitive data resides 'at rest' (question ***S1***) and how it is passed between parties is a key risk factor. Each system on which data is stored or processed is a point of potential breach or disclosure. Centralisation reduces the 'attack surface' of the ecosystem and may yield a more tractable security challenge, but increases the potential severity of a single breach. Data sensitivity also varies depending on aggregation and its processing history—this must be taken into account when answering question ***S2***. Processing sometimes reduces sensitivity, for example where unimportant data is not loaded or transferred, or where computation is information-destructive in other ways; in these cases 'intermediate' data is less sensitive than 'raw'. Furthermore, explicit disclosure control techniques such as sanitisation, aggregation, or obfuscation [36, 45] reduce sensitivity (these might be choices made when answering ***G2***). This means that in Mixed architectures we may see a stepwise decrease in data's sensitivity at each Processing activity. However, centralisation introduces sensitivity risk from data linkage attacks [100] in which multiple data sources, when combined, can yield new insights that were not previously apparent. These risks are known to be difficult to appraise a priori and so care must be taken to consider potential sources of related information available to particular recipients.

Costs. An well-known economic and fairness issue around centralisation is the allocation of costs, and the power dynamics that arise [14]. In terms of intermediaries this directly factors into questions ***S1*** and ***S2***: more centralised architectures impose greater data storage and processing costs on the intermediary; while decentralised architectures impose processing costs on suppliers that are likely to be greater in sum than the cost of the equivalent centralised processing task, as well as introducing overheads in time and bandwidth, and shifting energy costs to the supplier.

Software engineering. Centralised architectures (**Centralised** and **Mixed data-on-demand**) will be amenable to most programming paradigms, whereas distributed processing imposes additional constraints and decisions on the software engineer. 'Distributed systems' is a substantial field [19, 69], but existing paradigms may not map cleanly onto the roles in an intermediary system, especially in the case of secure multiparty computation [76], where stronger assurances can entail reduced functionality. This means that decisions made in response to structural specification questions (***S1–4***) might limit options for governance strategies (questions ***G1–4***). Novel combinations of governance tools and distributed computing are likely to yield substantial open problems.

4.4.2 Inter-party dynamics.

Supplier heterogeneity. Heterogeneity among suppliers' technical systems, interests, and risk models must be taken into account when deciding structural specification (questions ***S1–4***), and strategies might be employed to manage them (specifically ***G3*** and ***G4***). Heterogeneity in technical systems might have benefits for robustness and reliability in the face of attacks or outages. Similarly, having some better-resourced suppliers may enable load balancing under some distributed processing models [69]. However, differences in technical stacks pose engineering challenges for deployment and monitoring. Heterogeneity in risk appetite or privacy preferences might require accommodation—processing [19] can enable some suppliers to withhold data or only partly engage with a task (particularly relevant to ***G3***). Heterogeneity in interests also requires decision-making to set policy (which must be tackled in ***G4***); this must be done *a priori* in **Centralised** architectures, but in others policy may be set in real-time, e.g. with consensus algorithms.

Adversarial/competitive behaviour and trust. The threat of adversarial recipients (or suppliers) may motivate reliance on the intermediary as a well-resourced and protected single point of contact for all other parties, reducing the severity of any breach or attack. Adversarial intermediaries can cause significant risks, since they are by definition at least partly trusted even in **Conductor** architectures where there is a risk it can violate suppliers' interests through well-crafted requests. Adversarial suppliers may introduce risk for other suppliers, even in **Centralised** architectures, such as through data poisoning attacks [121]. Distributed processing

[4]Our implementation consisted of generic processing tasks to validate our architectures and inform our analysis. Measurements or quantitative evaluation of an intermediary system are only meaningful within the context of deployment and its tradeoffs, hence we do not present implementation details, as our focus is on the fundamental concepts underlying intermediary design.

models are usually devised with tolerance [69] for non-compliant or adversarial participants. Suppliers may not be malicious but be competitors, e.g. where the intermediary is managing a market. In these cases suppliers' trust in intermediary will affect the degree of centralisation adopted. Mechanisms of establishing trust between parties are therefore needed for an intermediary ecosystem to be viable—perhaps at the outset as part of ***S4***, or actively applied by governance strategies (***G1–4***). This can be achieved through transparency, or might be enforced by imposing costs on adversarial behaviour via contracts or fiduciary relationships. The EU DGA [96] attempts to set a baseline of trust by imposing obligations and costs on certain forms of intermediary, while data protection regulations do so on any parties handling personal data.

4.4.3 Mechanisms of governance. Every activity during the performance an intermediated task presents a point at which a data governance tool can be applied; the distinct 'surface' presented by each architecture thus influences the governance strategies that may be developed (i.e. the outputs of ***G1–4***).

Direct intervention. Strategies for ***G2*** and ***G3*** may require intervention during the processing task. The supplier is unable to actively intervene in **Centralised**, and in **Conductor** the intermediary's only opportunity is when instructing suppliers (**QIS**). In **Mixed**, both intermediary and suppliers may intervene, allowing for tradeoffs where each supplier addresses their own risk appetite while the intermediary ensures overall functionality.

Visibility and audit. Activities provide opportunities for parties to exercise visibility (i.e. ***G1*** strategies)—e.g. the intermediary (except in **Conductor**) can monitor recipient requests and information flows to maintain a complete picture of suppliers' per-recipient risk exposure. Similarly, other than in **Centralised**, suppliers can monitor in real time which data of theirs is being used—alongside technical disclosure budgets such as in differential privacy, they may moderate or prevent disclosure past certain limits. Record-keeping is often also required by regulation [22], and so parties may employ audit tools for compliance, especially when multiple sources of records (i.e. suppliers and intermediaries) are available to give confidence in accurate reporting. On the other hand, there is a risk of record fragmentation which might be mitigated via the intermediary under certain models for ensuring data consistency [30]. In some fields [44], data provenance [62] is fundamental to to ensuring data quality [113].

Secure multiparty computation and Privacy-Enhancing Technologies (PETs). Secure multiparty computation (MPC) algorithms [76] have models of inter-party trust built into their design, directly addressing ***G2*** and ***G3***. These computational approaches often incorporate a trusted (or partly-trusted) central party, and so map directly onto **Mixed** architectures. For example, consensus algorithms often designate a single aggregator for votes [71, 91], and federated learning algorithms have participants (suppliers) contributing incrementally to a model held by the central party (the intermediary) [70]. Other MPC algorithms proceed in an entirely peer-to-peer fashion after the task is initiated, thus requiring no work from the intermediary other than code generation, so will be applicable to **Conductor** architectures. PETs can be applied during processing to limit unwanted disclosures that might result from release. For example, in a **Mixed data-on-demand** architecture a supplier might apply sanitising or privacy-preserving transformations to data before releasing the data to the intermediary, so the intermediary never has access to sensitive information.

Legal relationships. Since architecture choice assigns roles to the parties, it will overlap with legal framings of rights, responsibilities, and obligations across multiple legal spheres, such as controllers and processors under GDPR [97], controls and remedies for breach in trade secrets law [95], and the EU DGA [96] which explicitly tackles intermediaries. Contractual means may also be employed to actively support governance strategies (***G1–G4***).

5 Connections to technical research

Establishing a technical study of intermediaries is particularly timely, as the data management community increasingly identifies governance as a central issue [1, 31, 86, 119]. Since work in data governance is framed using socio-legal language, and system requirements and operation are discussed with reference to law and compliance, identifying common problems is a vital step towards bridging perspectives with technical communities.

In order to join up the data governance literature with technical research fields, we identify existing work in data management and data security which are applicable to the implementation of data intermediaries. No one body of technical work is a perfect fit, so intermediaries are sure to pose a rich set of research questions and opportunities for novel technical solutions. We therefore also present a forward-looking technical agenda (§5.2), highlighting relevant trends and open problems in data management.

5.1 Relevant existing technical work

Some technical approaches from the fields of data management and distributed processing might be useful for operationalising intermediaries, either directly or as a starting point: **Centralised** and **Mixed data-on-demand** architectures can benefit from the *data integration* literature [31, 33, 86], while **Conductor** and **Mixed** architectures have clear parallels in multi-party processing. As a starting point for potential practitioners, we highlight in bold the key terms used for relevant technologies as well as relevant design questions from Table 3.

5.1.1 Centralised (and Mixed data on-demand). OnLine Analytical Processing (**OLAP**) techniques from data management [125, 126] will be applicable to Centralised or mixed data-on-demancd architectures which provide recipients with a hosted interface for processing. This would enable oversight and accountability (***G1***,***G4***) and opportunities to enforce rules and boundaries (***G2***).

Materialised data integration systems [31], which store data from multiple sources in a single system, present potential implementation tools for the more centralised architectures. The particular approach depends on the required distribution of data and the expected dynamics of its handling (***S1***). The **data warehouses** adopt *schema-on-write*—meaning data is integrated and transformed before use (i.e. before **RSI**)—which would be appropriate for purposes of pooling data or enabling collective action (see Table 2), *if* data and consent are unlikely to be withdrawn, *and if* the same policies

apply to all suppliers. However, for cases where supplier data or policy are heterogeneous, or data might be withdrawn, **data lakes'** [1] *schema-on-read* model performs minimal transformation of supplier data prior to the processing task. For particularly diverse data (see §5.2.2) Peer Data Management Systems (**PDMS**) [13, 54, 55] might provide useful tools.

5.1.2 Mixed and Conductor. Mixed intermediaries have some data on hand or results pre-computed to fulfil simple or common recipient requests, and query suppliers when additional data is needed (***S1***, ***S2***, ***S4***). This is the same data management approaches as **mediators** [131] and **virtual integration** systems with caches [47, 137]. These systems often harmonise heterogeneous data into a global schema [72, 74] or ontology [99].

Mixed (and Conductor) architectures also resemble processing models in Secure Multi-Party Computation (**SMPC**) [76], which use technical guarantees to secure distributed computing tasks, and which reflect or enforce particular models of inter-party trust (***G2*** and ***G3***). Some SMPC approaches rely on a trusted third party, which is applicable to certain Mixed use-cases, such as aggregating votes from suppliers with different preferences (***G3***) [71, 91].

Where confidentiality of supplier data is a primary concern (***G2***), a number of Privacy-Enhancing Technologies (**PETs**) offer relevant tools: **Homomorphic Encryption** allows computation on data without decrypting it first [34]; **Trusted Execution Environments** provide sealed environments where pre-agreed **PI** programs can be executed without interference from an untrusted intermediary [106, 134]; PETs for 'untrusted curator' threats such as **randomised response** and **local differential privacy** [38] modify supplier data at **PS**, preventing the disclosure of certain information to the intermediary or supplier, in ways that still enable useful aggregate analysis; and **federated learning** allows suppliers to train a shared model held at the intermediary without divulging their raw data to the intermediary or other suppliers [11, 70].

5.2 Technical research trends and challenges

5.2.1 Exploratory data science on distributed data. The data management community is already alert to the challenge of enabling data exploration of large heterogeneous datasets [1, 86]. While approaches from data lakes can be a starting point, limiting the recipient's access to data poses practical and privacy complications, for which work in Federated Learning may contribute solutions [11, 133]

5.2.2 Schema diversity. Data discovery in distributed systems (e.g. mixed aggregator/true-split architectures) encounters difficulties in building a global schema where datasets are large, diverse, change quickly, or unstructured [13, 50]. This will be especially true for intermediary ecosystems whose suppliers are not highly technical, where novel data uses cannot be anticipated, or where the intermediary must be agile in adapting to societal, legal, or other pressures. There has been work in data management on flexible or updateable schemas, such as Peer Data Management Systems (PDMS) [13, 54, 55], but this remains an open area of study.

5.2.3 Data discovery over query discovery. A recent trend in data integration towards 'data discovery' over 'query discovery', meaning a focus on exploring data rather than handling its structure (see Miller [82] for definitions and Nargesian et al. [86] for a survey of approaches), may allow new systems to be built with data minimisation in mind from the outset.

5.2.4 Adapting and leveraging software ecosystems. One of the most notable recent trends in data management is the disaggregation of compute and storage through cloud and orchestrated data management systems [1, 75]. This, along with a shift in data science towards more interactive notebook-style analysis, has resulted in a significant tooling ecosystem that supports various forms of distribution of computation and data storage.

5.2.5 Privacy-preserving data integration and confidential data management. While in §5.1 we noted the mature data integration literature that can support structural aspects of intermediaries (questions ***S1–4***), and governance tools from SMPC and PETs that can implement governance aims (***G1–G4***), combining the two is difficult. Privacy-preserving data integration (PPDI) is one area of crossover, producing frameworks that combine data integration and disclosure control techniques [8, 12, 21, 112]; Confidential Cloud Computing [105, 134] is another area, which focuses on combining distributed and delegated computing architectures with disclosure control. These approaches are useful but not sufficient for the governance tasks of intermediaries, since they focus mainly on data subject privacy; intermediaries may consider wider governance issues such as commercially sensitive data [107]. Some PPDI work has tackled wider notions of governance, such as one work which enacts purpose limitation by technical means [135]. A relevant open problem in the PET community is constructing privacy guarantees or metrics [129] and ensuring that these are meaningful to subjects [29] and practitioners [35]—such as interpreting the privacy parameter ϵ in differential privacy [84, 108]—and there is debate over whether any technical metric can capture the contextual nature of privacy [89].

5.2.6 Tension between audit/provenance/transparency and privacy. Data context and provenance are a significant concern in data management systems [1, 13, 62, 113, 117], just as they are for data intermediaries, specifically in record-keeping (***G1***), security and privacy (***G2***, ***G3***), and policy enforcement (***G2***) [20, 79, 88, 93, 94]. However, provenance information might disclose sensitive information about suppliers or their handling practices, which practitioners must be aware of—a known concern in data security [10, 16].

6 Conclusion

This paper is a first step towards bridging the gap between technical scholarship and practice and the growing data intermediaries community. Intermediaries have gained serious traction with scholars, governments, and regulators, and it is past time for their system design to be properly studied as a cross-disciplinary exercise, in order to support coming real-world deployment. Our work connects high-level governance goals and principles with practice, through our governance questions which allow practitioners to begin framing their task as a socio-technical system design problem, and our architecture-based language and generic API which provide a baseline for actually building and prototyping data intermediary ecosystems. We also provide a substantial initial study

of how system design impacts and enables governance, by elaborating a range of considerations for system designers. Finally, we explicitly identified the areas of technical research which need to be brought into the conversation around data intermediaries and data governance more generally. We hope this work forms a basis for a robust future cross-disciplinary practice in data governance.

Acknowledgments

We acknowledge support by the Engineering and Physical Sciences Research Council [grant number 1940975] and the Research Center Trustworthy Data Science and Security, University Alliance Ruhr.

References

[1] Daniel Abadi, Anastasia Ailamaki, David Andersen, Peter Bailis, Magdalena Balazinska, Philip Bernstein, Peter Boncz, Surajit Chaudhuri, Alvin Cheung, AnHai Doan, Luna Dong, Michael J. Franklin, Juliana Freire, Alon Halevy, Joseph M. Hellerstein, Stratos Idreos, Donald Kossmann, Tim Kraska, Sailesh Krishnamurthy, Volker Markl, Sergey Melnik, Tova Milo, C. Mohan, Thomas Neumann, Beng Chin Ooi, Fatma Ozcan, Jignesh Patel, Andrew Pavlo, Raluca Popa, Raghu Ramakrishnan, Christopher Ré, Michael Stonebraker, and Dan Suciu. 2020. The Seattle Report on Database Research. *ACM SIGMOD Record* 48, 4 (Feb. 2020), 44–53. doi:10.1145/3385658.3385668

[2] Environment Agency. 2024. Environment Agency Publishes Major Update to National Flood and Coastal Erosion Risk Assessment. https://www.gov.uk/government/news/environment-agency-publishes-major-update-to-national-flood-and-coastal-erosion-risk-assessment

[3] Ross Anderson. 2020. *Security Engineering: A Guide to Building Dependable Distributed Systems* (third edition ed.). John Wiley & Sons, Indianapolis.

[4] Theo Araujo, Jef Ausloos, Wouter van Atteveldt, Felicia Loecherbach, Judith Moeller, Jakob Ohme, Damian Trilling, Bob van de Velde, Claes De Vreese, and Kasper Welbers. 2022. OSD2F: An Open-Source Data Donation Framework. *Computational Communication Research* 4, 2 (2022), 372–387. doi:10.5117/CCR2022.2.001.ARAU

[5] Maurizio Atzori, Angelo Ciaramella, Claudia Diamantini, B. D. Martino, Salvatore Distefano, Tullio Facchinetti, Fabrizio Montecchiani, Antonino Nocera, Giancarlo Ruffo, and Roberto Trasarti. 2024. Dataspaces: Concepts, Architectures and Initiatives. In *Proceedings of the 2nd Italian Conference on Big Data and Data Science (ITADATA 2023) (ITADATA 2023, Vol. 3606)*. CEUR-WS, Naples, Italy, 1–14. https://iris.unica.it/handle/11584/389724

[6] Jef Ausloos and Michael Veale. 2021. Researching with Data Rights. *Technology and Regulation* 2020 (Jan. 2021), 136–157. doi:10.26116/techreg.2020.010

[7] David Beer. 2017. The Social Power of Algorithms. *Information, Communication & Society* 20, 1 (2017), 1–13. doi:10.1080/1369118X.2016.1216147

[8] Djamal Benslimane, Mahmoud Barhamgi, Frederic Cuppens, Franck Morvan, Bruno Defude, Ebrahim Nageba, Michael Mrissa, Francois Paulus, Stephane Morucci, Nora Cuppens, Chirine Ghedira, Riad Mokadem, Said Oulmakhzoune, and Jocelyne Fayn. 2013. PAIRSE: A Privacy-Preserving Service-Oriented Data Integration System. *ACM SIGMOD Record* 42, 3 (Oct. 2013), 42–47. doi:10.1145/2536669.2536677

[9] Sebastian Benthall and Jake Goldenfein. 2021. Artificial Intelligence and the Purpose of Social Systems. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. ACM, Virtual Event USA, 3–12. doi:10.1145/3461702.3462526

[10] Elisa Bertino, Gabriel Ghinita, Murat Kantarcioglu, Dang Nguyen, Jae Park, Ravi Sandhu, Salmin Sultana, Bhavani Thuraisingham, and Shouhuai Xu. 2014. A Roadmap for Privacy-Enhanced Secure Data Provenance. *Journal of Intelligent Information Systems* 43, 3 (Dec. 2014), 481–501. doi:10.1007/s10844-014-0322-7

[11] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D. Lane. 2022. Flower: A Friendly Federated Learning Research Framework. doi:10.48550/arXiv.2007.14390 arXiv:2007.14390 [cs]

[12] S.S. Bhowmick, L. Gruenwald, M. Iwaihara, and S. Chatvichienchai. 2006. PRIVATE-IYE: A Framework for Privacy Preserving Data Integration. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, Atlanta, GA, USA, 91–91. doi:10.1109/ICDEW.2006.117

[13] Tobias Bleifuß, Leon Bornemann, Theodore Johnson, Dmitri V. Kalashnikov, Felix Naumann, and Divesh Srivastava. 2018. Exploring Change: A New Dimension of Data Analytics. *Proceedings of the VLDB Endowment* 12, 2 (Oct. 2018), 85–98. doi:10.14778/3282495.3282496

[14] Balázs Bodó, Jaya Klara Brekke, and Jaap-Henk Hoepman. 2021. Decentralisation: A Multidisciplinary Perspective. *Internet Policy Review* 10, 2 (2021), 1–21. doi:10.14763/2021.2.1563

[15] Balázs Bodó, Kristina Irion, Heleen Janssen, and Alexandra Giannopoulou. 2021. Personal Data Ordering in Context: The Interaction of Meso-Level Data Governance Regimes with Macro Frameworks. *Internet Policy Review* 10, 3 (Sept. 2021), 1–31. doi:10.14763/2021.3.1581

[16] Uri Braun, Avraham Shinnar, and Margo Seltzer. 2008. Securing Provenance. In *Proceedings of the 3rd Conference on Hot Topics in Security (HOTSEC'08)*. USENIX Association, USA, 1–5.

[17] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[18] M Bunting and S Lansdell. 2019. Designing Decision Making Processes for Data Trusts: Lessons from Three Pilots. https://www.involve.org.uk/sites/default/files/field/attachemnt/General-decision-making-report-Apr-19.pdf

[19] Brendan Burns. 2018. *Designing Distributed Systems: Patterns and Paradigms for Scalable, Reliable Services* (first edition ed.). O'Reilly, Beijing Boston.

[20] James Cheney. 2011. A Formal Framework for Provenance Security. In *2011 IEEE 24th Computer Security Foundations Symposium*. IEEE, Cernay-la-Ville, France, 281–293. doi:10.1109/CSF.2011.26

[21] Chris Clifton, Murat Kantarcioğlu, AnHai Doan, Gunther Schadow, Jaideep Vaidya, Ahmed Elmagarmid, and Dan Suciu. 2004. Privacy-Preserving Data Integration and Sharing. In *Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD '04)*. Association for Computing Machinery, New York, NY, USA, 19–26. doi:10.1145/1008694.1008698

[22] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 598–609. doi:10.1145/3442188.3445921

[23] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding Accountability in Algorithmic Supply Chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1186–1197. doi:10.1145/3593013.3594073

[24] European Commission. 2024. *Staff Working Document on Common European Data Spaces*. Technical Report. European Commission. https://digital-strategy.ec.europa.eu/en/library/second-staff-working-document-data-spaces

[25] European Commssion. 2020. A European Strategy for Data. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066&from=ENhttps://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066&from=EN

[26] UK AI Council and Ada Lovelace Institute. 2021. Exploring Legal Mechanisms for Data Stewardship. Retrieved April 13, 2022 from https://www.adalovelaceinstitute.org/report/legal-mechanisms-data-stewardship/

[27] Andy Crabtree, Tom Lodge, James Colley, Chris Greenhalgh, Kevin Glover, Hamed Haddadi, Yousef Amar, Richard Mortier, Qi Li, John Moore, Liang Wang, Poonam Yadav, Jianxin Zhao, Anthony Brown, Lachlan Urquhart, and Derek McAuley. 2018. Building Accountability into the Internet of Things: The IoT Databox Model. *Journal of Reliable Intelligent Environments* 4, 1 (April 2018), 39–55. doi:10.1007/s40860-018-0054-5

[28] Ella Creamer. 2025. US Authors' Copyright Lawsuits against OpenAI and Microsoft Combined in New York with Newspaper Actions. https://www.theguardian.com/books/2025/apr/04/us-authors-copyright-lawsuits-against-openai-and-microsoft-combined-in-new-york-with-newspaper-actions

[29] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. 2021. "I Need a Better Description": An Investigation Into User Expectations For Differential Privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)*. Association for Computing Machinery, New York, NY, USA, 3037–3052. doi:10.1145/3460120.3485252

[30] Susan B Davidson, Hector Garcia-Molina, and Dale Skeen. 1985. Consistency in a Partitioned Network: A Survey. *ACM Computing Surveys (CSUR)* 17, 3 (1985), 341–370.

[31] Ali Davoudian and Mengchi Liu. 2020. Big Data Systems: A Software Engineering Perspective. *Comput. Surveys* 53, 5 (Sept. 2020), 110:1–110:39. doi:10.1145/3408314

[32] Sylvie Delacroix and Neil D Lawrence. 2019. Bottom-up Data Trusts: Disturbing the 'One Size Fits All' Approach to Data Governance. *International Data Privacy Law* 9, 4 (Nov. 2019), 236–252. doi:10.1093/idpl/ipz014

[33] AnHai Doan. 2012. *Principles of Data Integration*. Morgan Kaufmann, Waltham, MA. https://books.google.com/books?hl=en&lr=&id=s2YCKGrO10YC&oi=fnd&pg=PP2&dq=AnHai+Doan,+Alon+Halevy,+and+Zachary+Ives.+2012.+Principles+of+Data+Integration.+Morgan+Kaufmann.&ots=SPMJuZNYT9&sig=A1a4pV8CmjaNN9uKUXa-K6WaXaw

[34] Thi Van Thao Doan, Mohamed-Lamine Messai, Gérald Gavin, and Jérôme Darmont. 2023. A Survey on Implementations of Homomorphic Encryption Schemes. *The Journal of Supercomputing* 79, 13 (2023), 15098–15139.

[35] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. 2019. Differential Privacy in Practice: Expose Your Epsilons! *Journal of Privacy and Confidentiality* 9, 2 (Oct. 2019), 1–22. doi:10.29012/jpc.689

[36] Cynthia Dwork and Moni Naor. 2010. On the Difficulties of Disclosure Prevention in Statistical Databases or The Case for Differential Privacy. *Journal of Privacy and Confidentiality* 2, 1 (Sept. 2010), 93–107. doi:10.29012/jpc.v2i1.585

[37] Lilian Edwards. 2005. The Problem with Privacy–A Modest Proposal.

[38] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. Association for Computing Machinery, New York, NY, USA, 1054–1067. doi:10.1145/2660267.2660348

[39] Centre for Data Ethics and Innovation. 2021. Unlocking the Value of Data: Exploring the Role of Data Intermediaries. https://www.gov.uk/government/publications/unlocking-the-value-of-data-exploring-the-role-of-data-intermediaries

[40] Worker Info Exchange. 2021. Worker Info Exchange Research. https://www.workerinfoexchange.org/research

[41] Eimear Farrell, Marco Minghini, Alexander Kotsev, Josep Soler Garrido, Brooke Tapsall, Marina Micheli, Monica Posada Sanchez, Serena Signorelli, Alessio Tartaro, Jaime Bernal Cereceda, Michele Vespe, Margherita Di Leo, Bruno Carballa Smichowski, Robin Smith, Sven Schade, Katarzyna Pogorzelska, Lorenzo Gabrielli, and Davide de Marchi. 2023. *European Data Spaces - Scientific Insights into Data Sharing and Utilisation at Scale.* Technical Report JRC129900. Joint Research Centre. doi:10.2760/400188

[42] Open Humans Foundation. 2024. Open Humans. https://www.openhumans.org/

[43] E. Frantz. 2019. Q and A: Fighting for Worker's Right to Data. https://www.opensocietyfoundations.org/voices/q-and-a-fighting-for-workers-right-to-data

[44] Xin Fu, Anna Wojak, Daniel Neagu, Mick Ridley, and Kim Travis. 2011. Data Governance in Predictive Toxicology: A Review. *Journal of Cheminformatics* 3 (2011), 1–16. doi:10.1186/1758-2946-3-24

[45] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *Comput. Surveys* 42, 4 (June 2010), 14:1–14:53. doi:10.1145/1749603.1749605

[46] Erich Gamma (Ed.). 2011. *Design Patterns: Elements of Reusable Object-Oriented Software* (39. printing ed.). Addison-Wesley, Boston, Mass. Munich.

[47] Felix Gessert, Michael Schaarschmidt, Wolfram Wingerath, Erik Witt, Eiko Yoneki, and Norbert Ritter. 2017. Quaestor: Query Web Caching for Database-as-a-Service Providers. *Proceedings of the VLDB Endowment* 10, 12 (Aug. 2017), 1670–1681. doi:10.14778/3137765.3137773

[48] Global Partnership on AI (GPAI), Aapti Institute, and Open Data Institute. 2021. *Enabling Data Sharing for Social Benefit through Data Trusts. An Interim Report for the 2021 GPAI Paris Summit.* Technical Report. Global Partnership on AI (GPAI). https://gpai.ai/projects/data-governance/data-trusts/

[49] Global Partnership on AI (GPAI), Aapti Institute and Open Data Institute. 2021. Enabling Data Sharing for Social Benefit through Data Trusts. An Interim Report for the 2021 GPAI Paris Summit. https://gpai.ai/projects/data-governance/data-trusts/

[50] Behzad Golshan, Alon Halevy, George Mihaila, and Wang-Chiew Tan. 2017. Data Integration: After the Teenage Years. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS '17)*. Association for Computing Machinery, New York, NY, USA, 101–106. doi:10.1145/3034786.3056124

[51] UK Government. 2021. Check the Long Term Flood Risk for an Area in England. https://www.gov.uk/check-long-term-flood-risk

[52] Karen Gregory. 2021. 'Worker Data Science' Can Teach Us How to Fix the Gig Economy. https://www.wired.com/story/labor-organizing-unions-worker-algorithms/

[53] The W3C Solid Community Group. 2016. The Solid Project. https://solidproject.org/

[54] A.Y. Halevy, Z.G. Ives, J. Madhavan, P. Mork, D. Suciu, and I. Tatarinov. 2004. The Piazza Peer Data Management System. *IEEE Transactions on Knowledge and Data Engineering* 16, 07 (July 2004), 787–798. doi:10.1109/TKDE.2004.1318562

[55] Alon Y. Halevy, Zachary G. Ives, Dan Suciu, and Igor Tatarinov. 2005. Schema Mediation for Large-Scale Semantic Data Sharing. *The VLDB Journal* 14, 1 (March 2005), 68–83. doi:10.1007/s00778-003-0116-y

[56] Wendy Hall and Jérôme Pesenti. 2017. *Growing the Artificial Intelligence Industry in the UK.* Technical Report. Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy.

[57] Jack Hardinges and Jared Robert Keller. 2021. What Are 'bottom-up' Data Institutions and How Do They Empower People? https://theodi.org/news-and-events/blog/what-are-bottom-up-data-institutions-and-how-do-they-empower-people

[58] Jack Hardinges and Jared Robert Keller. 2021. What Are Data Institutions and Why Are They Important? https://theodi.org/article/what-are-data-institutions-and-why-are-they-important/

[59] Jack Hardinges, Peter Wells, Alex Blandford, Jeni Tennison, and Anna Scott. 2019. *Data Trusts: Lessons from Three Pilots.* Technical Report. Open Data Institute. https://theodi.org/news-and-events/blog/odi-data-trusts-report/

[60] Thomas Hardjono and Alex Pentland. 2019. Data Cooperatives: Towards a Foundation for Decentralized Personal Data Management. doi:10.48550/arXiv.1905.08819 arXiv:1905.08819

[61] Todd Hartman, Helen Kennedy, Robin Steedman, and Rhianne Jones. 2020. Public Perceptions of Good Data Management: Findings from a UK-based Survey. *Big Data & Society* 7, 1 (2020), 2053951720935616. doi:10.1177/2053951720935616

[62] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. 2017. A Survey on Provenance: What for? What Form? What From? *Vldb Journal* 26, 6 (2017), 881–906. doi:10.1007/S00778-017-0486-1

[63] Saffron Huang and Divya Siddarth. 2023. Generative AI and the Digital Commons. doi:10.48550/arXiv.2303.11074 arXiv:2303.11074 [cs]

[64] Heleen Janssen, Jennifer Cobbe, Chris Norval, and Jatinder Singh. 2020. Decentralized Data Processing: Personal Data Stores and the GDPR. *International Data Privacy Law* 10, 4 (Nov. 2020), 356–384. doi:10.1093/idpl/ipaa016

[65] Heleen Janssen, Jennifer Cobbe, and Jatinder Singh. 2020. Personal Information Management Systems: A User-Centric Privacy Utopia? *Internet Policy Review* 9, 4 (Dec. 2020), 1–25. doi:10.14763/2020.4.1536

[66] Heleen Janssen and Jatinder Singh. 2022. Data Intermediary. *Internet Policy Review* 11, 1 (2022), 9. doi:10.14763/2022.1.1644

[67] Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2206–2222. doi:10.1145/3531146.3534637 arXiv:2206.03216 [cs]

[68] Rob Kitchin. 2017. Thinking Critically about and Researching Algorithms. *Information, communication & society* 20, 1 (2017), 14–29.

[69] Martin Kleppmann. 2017. *Designing Data-Intensive Applications: The Big Ideas behind Reliable, Scalable, and Maintainable Systems* (first edition ed.). O'Reilly, Beijing Boston Farnham Sebastopol Tokyo.

[70] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. doi:10.48550/arXiv.1610.02527 arXiv:1610.02527

[71] Leslie Lamport. 2001. Paxos Made Simple. *ACM SIGACT News (Distributed Computing Column) 32, 4 (Whole Number 121, December 2001)* 32, 4 (Dec. 2001), 51–58. doi:10.1145/568425.568433

[72] Andreas Langegger, Wolfram Wöß, and Martin Blöchl. 2008. A Semantic Web Middleware for Virtual Data Integration on the Web. In *The Semantic Web: Research and Applications*, Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis (Eds.). Springer, Berlin, Heidelberg, 493–507. doi:10.1007/978-3-540-68234-9_37

[73] Tuukka Lehtiniemi. 2017. Personal Data Spaces: An Intervention in Surveillance Capitalism? *Surveillance & Society* 15, 5 (2017), 626–639. doi:10.24908/ss.v15i5.6424

[74] Maurizio Lenzerini. 2002. Data Integration: A Theoretical Perspective. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '02)*. Association for Computing Machinery, New York, NY, USA, 233–246. doi:10.1145/543613.543644

[75] Feifei Li. 2023. Modernization of Databases in the Cloud Era: Building Databases That Run Like Legos. *Proceedings of the VLDB Endowment* 16, 12 (Aug. 2023), 4140–4151. doi:10.14778/3611540.3611639

[76] Yehuda Lindell. 2020. Secure Multiparty Computation. *Commun. ACM* 64, 1 (2020), 86–96. doi:10.1145/3387108

[77] René LP Mahieu and Jef Ausloos. 2020. Harnessing the Collective Potential of GDPR Access Rights: Towards an Ecology of Transparency. *Internet Policy Review* 6 (2020), 1.

[78] Marcello M. Mariani, Maria Ek Styven, and Fréderic Teulon. 2021. Explaining the Intention to Use Digital Personal Data Stores: An Empirical Study. *Technological Forecasting and Social Change* 166 (May 2021), 120657. doi:10.1016/j.techfore.2021.120657

[79] Andrew Martin, John Lyle, and Cornelius Namilkuo. 2012. Provenance as a Security Control. In *Proceedings of the 4th USENIX Conference on Theory and Practice of Provenance (TaPP'12)*. USENIX Association, USA, 3. https://www.usenix.org/conference/tapp12/workshop-program/presentation/martin

[80] Marina Micheli, Eimear Farrell, SMICHOWSKI Bruno Carballa, SANCHEZ Monica Posada, Serena Signorelli, and Michele Vespe. 2023. Mapping the Landscape of Data Intermediaries. https://www.researchgate.net/profile/Serena-Signorelli/publication/375888375_Mapping_the_landscape_of_data_intermediaries/links/6560ae70b1398a779dad3e02/Mapping-the-landscape-of-data-intermediaries.pdf

[81] Marina Micheli, Marisa Ponti, Max Craglia, and Anna Berti Suman. 2020. Emerging Models of Data Governance in the Age of Datafication. *Big Data & Society*

7, 2 (July 2020), 2053951720948087. doi:10.1177/2053951720948087
[82] Renée J. Miller. 2018. Open Data Integration. *Proceedings of the VLDB Endowment* 11, 12 (Aug. 2018), 2130–2139. doi:10.14778/3229863.3240491
[83] Fernando Monge, Sarah Barns, Rainer Kattel, and Francesca Bria. 2022. *A New Data Deal: The Case of Barcelona.* Working / Discussion Paper 2022/02. UCL Institute for Innovation and Public Purpose, London, UK. https://www.ucl.ac.uk/bartlett/public-purpose/publications/2022/feb/new-data-deal-case-barcelona
[84] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. 2023. What Are the Chances? Explaining the Epsilon Parameter in Differential Privacy. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 1613–1630. https://www.usenix.org/conference/usenixsecurity23/presentation/nanayakkara
[85] Arvind Narayanan, Solon Barocas, Vincent Toubiana, Helen Nissenbaum, and Dan Boneh. 2012. A Critical Look at Decentralized Personal Data Architectures. *Network (Bristol, England)* 25 (2012), 8.
[86] Fatemeh Nargesian, Erkang Zhu, Renée J. Miller, Ken Q. Pu, and Patricia C. Arocena. 2019. Data Lake Management: Challenges and Opportunities. *Proceedings of the VLDB Endowment* 12, 12 (Aug. 2019), 1986–1989. doi:10.14778/3352063.3352116
[87] I. Ng and H Haddadi. 2018. Decentralised AI Has the Potential to Upend the Online Economy. https://www.wired.co.uk/article/decentralised-artificial-intelligence
[88] Qun Ni, Shouhuai Xu, Elisa Bertino, Ravi Sandhu, and Weili Han. 2009. An Access Control Language for a General Provenance Model. In *Secure Data Management*, Willem Jonker and Milan Petković (Eds.). Springer, Berlin, Heidelberg, 68–88. doi:10.1007/978-3-642-04219-5_5
[89] Helen Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life.* Stanford University Press, Stanford, CA. doi:10.1515/9780804772891
[90] Kieron O'Hara. 2019. *Data Trusts: Ethics, Architecture and Governance for Trustworthy Data Stewardship.* WSI White Paper 1. Web Science Institute, Southampton, UK.
[91] Diego Ongaro and John Ousterhout. 2014. In Search of an Understandable Consensus Algorithm. In *USENIX ATC'14: Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference*. USENIX Association, Philadelphia, Pennsylvania, 305–319. https://www.usenix.org/conference/atc14/technical-sessions/presentation/ongaro
[92] Elinor Ostrom. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action* (1 ed.). Cambridge University Press, Cambridge. https://www.cambridge.org/core/product/identifier/9781316423936/type/book
[93] Bofeng Pan, Natalia Stakhanova, and Suprio Ray. 2023. Data Provenance in Security and Privacy. *Comput. Surveys* 55, 14s (July 2023), 323:1–323:35. doi:10.1145/3593294
[94] Jaehong Park, Dang Nguyen, and Ravi Sandhu. 2012. A Provenance-Based Access Control Model. In *2012 Tenth Annual International Conference on Privacy, Security and Trust*. IEEE, Paris, France, 137–144. doi:10.1109/PST.2012.6297930
[95] European Parliament and European Council. 2016. Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the Protection of Undisclosed Know-How and Business Information (Trade Secrets) against Their Unlawful Acquisition, Use and Disclosure. *Official Journal of the European Union* 157 (June 2016), 18.
[96] European Parliament and European Council. 2022. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance and Amending Regulation 2018/1724 (Data Governance Act).
[97] European Parliament and 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.
[98] Davor Petreski and Marc Cheong. 2024. Data Cooperatives: A Conceptual Review. In *International Conference on Information Systems 2024*. AIS eLibrary, Bangkok, Thailand, 19. https://aisel.aisnet.org/icis2024/lit_review/lit_review/15/
[99] Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. 2008. Linking Data to Ontologies. *Journal on Data Semantics X* 4900 (2008), 133–173. doi:10.1007/978-3-540-77688-8_5
[100] Jovan Powar and Alastair R. Beresford. 2023. SoK: Managing Risks of Linkage Attacks on Data Privacy. *Proceedings on Privacy Enhancing Technologies* 2 (2023), 97–116. doi:10.56553/popets-2023-0043
[101] J. Powar and J. Hancock. 2023. The Limits to Modelling Dataflow Networks: A Conceptual Synthesis. In *International Conference on AI and the Digital Economy (CADE 2023)*, Vol. 2023. IET, Venice, Italy, 134–143. doi:10.1049/icp.2023.2597
[102] Privacy Enhancing Technologies Working group. 2019. *Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis.* The Royal Society / The Royal Society, London, UK. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Protecting-privacy-in-practice.pdf
[103] Privacy Enhancing Technologies Working group. 2023. *From Privacy to Partnership: The Role of Privacy Enhancing Technologies in Data Governance and Collaborative Analysis.* The Royal Society / The Royal Society, London, UK. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/From-Privacy-to-Partnership.pdf
[104] C Reed, BPE Solicitors, and Pinsent Masons. 2019. Data Trusts: Legal and Governance Considerations.
[105] Mark Russinovich, Manuel Costa, Cédric Fournet, David Chisnall, Antoine Delignat-Lavaud, Sylvan Clebsch, Kapil Vaswani, and Vikas Bhatia. 2021. Toward Confidential Cloud Computing. *Commun. ACM* 64, 6 (June 2021), 54–61. doi:10.1145/3453930
[106] 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. IEEE, Helsinki, Finland, 57–64. doi:10.1109/Trustcom.2015.357
[107] Teresa Scassa. 2020. Designing Data Governance for Data Sharing: Lessons from Sidewalk Toronto. *Technology and Regulation* 2020 (July 2020), 44–56. doi:10.71265/d7yvsg86
[108] Jeremy Seeman and Daniel Susser. 2024. Between Privacy and Utility: On Differential Privacy in Theory and Practice. *ACM Journal on Responsible Computing* 1, 1 (March 2024), 3:1–3:18. doi:10.1145/3626494
[109] Ashraf Shaharudin, Bastiaan Van Loenen, and Marijn Janssen. 2023. Towards a Common Definition of Open Data Intermediaries. *Digital Government: Research and Practice* 4, 2 (June 2023), 1–21. doi:10.1145/3585537
[110] Ashraf Shaharudin, Bastiaan van Loenen, and MFWHA Janssen. 2023. Identifying Business Models of Open Data Intermediaries: A Review. In *International Conference on Open Data: Open Data Challenges and Opportunities in Times of Crisis and Growth (ICOD 2022, Vol. 1)*. Faculty of Law, University of Zagreb, Zagreb, Croatia, 13–18. https://research.tudelft.nl/en/publications/identifying-business-models-of-open-data-intermediaries-a-review
[111] Mary Shaw and David Garlan. 1996. *Software Architecture: Perspectives on an Emerging Discipline.* Prentice Hall, Upper Saddle River, NJ.
[112] Vijay Maruti Shelake and Narendra Shekokar. 2017. A Survey of Privacy Preserving Data Integration. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*. IEEE, Mysuru, 59–70. doi:10.1109/ICEECCOT.2017.8284559
[113] Jatinder Singh, Jennifer Cobbe, and Chris Norval. 2018. Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE access : practical innovations, open solutions* 7 (2018), 6562–6574.
[114] Veronika Siska, Vasileios Karagiannis, and Mario Drobics. 2023. *Building a Dataspace: Technical Overview.* Technical Report. Gaia-X Hub Austria, Austria. https://www.gaia-x.at/wp-content/uploads/2023/04/WhitepaperGaiaX.pdf
[115] Sophie Stalla-Bourdillon, Gefion Thuermer, Johanna Walker, Laura Carmichael, and Elena Simperl. 2020. Data Protection by Design: Building the Foundations of Trustworthy Data Sharing. *Data & Policy* 2 (2020), e4. doi:10.1017/dap.2020.1
[116] Chris Stokel-Walker. 2024. Here's the Deal: AI Giants Get to Grab All Your Data Unless You Say They Can't. Fancy That? No, Neither Do I. https://www.theguardian.com/commentisfree/2024/oct/18/ai-systems-big-tech-data-ministers
[117] Isuru Suriarachchi and Beth Plale. 2016. Provenance as Essential Infrastructure for Data Lakes. In *Provenance and Annotation of Data and Processes*, Marta Mattoso and Boris Glavic (Eds.). Springer International Publishing, Cham, 178–182. doi:10.1007/978-3-319-40593-3_16
[118] Iryna Susha, Marijn Janssen, and Stefaan Verhulst. 2017. Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. AIS eLibrary, Hawaii, 2691–2700. doi:10.24251/HICSS.2017.325
[119] Ignacio G. Terrizzano, Peter M. Schwarz, Mary Roth, and John E. Colino. 2015. Data Wrangling: The Challenging Journey from the Wild to the Lake.. In *Proceedings of the Biennial Conference on Innovative Data Systems Research (CIDR) (CIDR '15)*. Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, USA, 9. http://cidrdb.org/cidr2015/Papers/CIDR15_Paper2.pdf
[120] Peter Tolmie, Andy Crabtree, Tom Rodden, James Colley, and Ewa Luger. 2016. "This Has to Be the Cats": Personal Data Legibility in Networked Sensing Systems. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 491–502. doi:10.1145/2818048.2819992
[121] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. 2020. Data Poisoning Attacks Against Federated Learning Systems. In *Computer Security – ESORICS 2020*, Liqun Chen, Ninghui Li, Kaitai Liang, and Steve Schneider (Eds.). Springer International Publishing, Cham, 480–501. doi:10.1007/978-3-030-58951-6_24
[122] Jennifer Tridgell and Jatinder Singh. 2025. 'Stewardship' as a Fair, Accountable and Transparent Model for Free and Open-Source Software Governance? Looking Beyond the EU's Cyber Resilience Act. In *2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens, Greece, 0–0. doi:10.1145/3715275.3732032

[123] Tom M. van Engers, Robert Meijer, Leon Gommans, and Leon Nieuwenhuis. 2016. Secure Digital Marketplaces Supporting Cross-Organizational Production Processes. In *NWO/STW Workshop ICT with Industry 2016*. DL4LD, Leiden, 15. https://dl4ld.nl/2017-02-24-Secure-Virtual-Digital-Market-Places.pdf

[124] Max Van Kleek and Kieron OHara. 2014. The Future of Social Is Personal: The Potential of the Personal Data Store. In *Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society*, Daniele Miorandi, Vincenzo Maltese, Michael Rovatsos, Anton Nijholt, and James Stewart (Eds.). Springer International Publishing, Cham, 125–158. doi:10.1007/978-3-319-08681-1_7

[125] Panos Vassiliadis, Patrick Marcel, and Stefano Rizzi. 2019. Beyond Roll-up's and Drill-down's: An Intentional Analytics Model to Reinvent OLAP. *Information Systems* 85 (Nov. 2019), 68–91. doi:10.1016/j.is.2019.03.011

[126] Panos Vassiliadis and Timos Sellis. 1999. A Survey of Logical Models for OLAP Databases. *SIGMOD Rec.* 28, 4 (Dec. 1999), 64–69. doi:10.1145/344816.344869

[127] Daniel T. Wagner, Andrew Rice, and Alastair R. Beresford. 2014. Device Analyzer: Large-Scale Mobile Data Collection. *ACM SIGMETRICS Performance Evaluation Review* 41, 4 (April 2014), 53–56. doi:10.1145/2627534.2627553

[128] Daniel T. Wagner, Andrew Rice, and Alastair R. Beresford. 2014. Device Analyzer: Understanding Smartphone Usage. In *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, Ivan Stojmenovic, Zixue Cheng, and Song Guo (Eds.). Vol. 131. Springer International Publishing, Cham, 195–208. doi:10.1007/978-3-319-11569-6_16

[129] Isabel Wagner and David Eckhoff. 2019. Technical Privacy Metrics: A Systematic Survey. *Comput. Surveys* 51, 3 (May 2019), 1–38. doi:10.1145/3168389 arXiv:1512.00327 [cs, math]

[130] Algorithm Watch and Open Knowledge Foundation Germany. 2019. OpenSCHUFA: The Campaign Is over, the Problems Remain. https://openschufa.de/english/

[131] G. Wiederhold. 1992. Mediators in the Architecture of Future Information Systems. *Computer* 25, 3 (March 1992), 38–49. doi:10.1109/2.121508

[132] Janis Wong, Tristan Henderson, and Kirstie Ball. 2022. Data Protection for the Common Good: Developing a Framework for a Data Protection-Focused Data Commons. *Data & Policy* 4 (Jan. 2022), e3. doi:10.1017/dap.2021.40

[133] Yuncheng Wu, Naili Xing, Gang Chen, Tien Tuan Anh Dinh, Zhaojing Luo, Beng Chin Ooi, Xiaokui Xiao, and Meihui Zhang. 2023. Falcon: A Privacy-Preserving and Interpretable Vertical Federated Learning System. *Proc. VLDB Endow.* 16, 10 (June 2023), 2471–2484. doi:10.14778/3603581.3603588

[134] Zhifeng Xiao and Yang Xiao. 2013. Security and Privacy in Cloud Computing. *IEEE Communications Surveys & Tutorials* 15, 2 (2013), 843–859. doi:10.1109/SURV.2012.060912.00182

[135] Stephen S. Yau and Yin Yin. 2008. A Privacy Preserving Repository for Data Integration across Data Sharing Services. *IEEE Transactions on Services Computing* 1, 3 (July 2008), 130–140. doi:10.1109/TSC.2008.14

[136] Edward Yourdon and Larry L. Constantine. 1979. *Structured Design: Fundamentals of a Discipline of Computer Program and Systems Design*. Prentice Hall, Englewood Cliffs, N.J.

[137] Victor Zakhary, Divyakant Agrawal, and Amr El Abbadi. 2017. Caching at the Web Scale. *Proceedings of the VLDB Endowment* 10, 12 (Aug. 2017), 2002–2005. doi:10.14778/3137765.3137831

[138] X. Zhou. 2024. *From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation*. Ph. D. Dissertation. University of Amsterdam, Amsterdam. https://dare.uva.nl/search?identifier=c54ef847-63f9-43f6-b324-1c34261990f0

[139] Shoshana Zuboff. 2015. Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of information technology* 30, 1 (2015), 75–89. doi:10.1057/jit.2015.5