Privacy and Security Requirements for a Digital Data Hub

Martin Gfeller
Swisscom
Zurich, Switzerland
martin.gfeller@swisscom.com

Thomas Hardjono
MIT Connection Science & Engineering
Boston MA
hardjono@mit.edu

***Abstract*—A Digital Data Hub provides Data Accounts wherein persons may store data that are collected in their interaction with organizations. The hub is extensible by contributed Data Apps that gather and process data and conduct business transactions. They may act as Agents assisting the users in their daily lives. Private-banking-grade Privacy and Security ensure that the power of this data collection cannot be misused. Schema standardization is key to ensure privacy and security.**

Keywords—data, data accounts, data ecosystem, hub, privacy, security

1 The Vision of a Digital Data Hub

Imagine all the forms you fill out, all the documents and data created when you interact with companies large or small, with the government, with other organizations, or with other individuals are captured and made available in your *personal data account*. Like bank accounts, data accounts permit a lifelong collection and saving of data.

Imagine you can share selected subsets of your data with organizations, companies, governments agencies to simplify your interactions with them.

Imagine you can share your data with agents – human or automated, to pursue your interest, and help make decisions on your behalf.

Imagine organizations can access the data you share with them through *Data Apps,* which they or their partners develop. Data Apps import, export, and transform data, gaining knowledge based on data they were granted access to. Data Apps are executed in a secure environment, which ensures that your privacy is respected.

The *Digital Data Hub (DDH)* provides data accounts for everybody, an ecosystem to share and process data. It enables organizations to contribute and profit from a *data economy.* It enables startups to provide solutions, without being suffocated by the need to gather all data first.

By providing data accounts controlled by individuals, it contributes to their digital sovereignty, and ultimately to the digital sovereignty of the country [16].

Individuals today are more aware of the value of their personal data to improve their lives and the lives of others in the community. However, they are also wary of the negative potential of unregulated collections of personal data [9][5].

The main aim of a DDH is to ease the exchange of information between individuals and organizations, and therefore to simplify their interactions, focussing on what is not straight-through digital today.

Data standards promote interoperability and avoid data lock-in with organizations that collect and store massive amounts of personal data. Implicit consent (by agreement to terms and conditions) and the legal right of data portability [17] are replaced by explicitly granted consent and data stewardship by the DDH.

The DDH is an organization and market for solutions based on data. It focusses on the multitude of small relationships an individual has; not on the big data an organization may assemble from many individuals.

The use cases are manifold, and only limited by imagination. Initially, we expect an emphasis on (digital) agents and brokers that simplify services for individuals. We except an evolution into a decentralized eco system of data providers, enhancers, and consumers, leading to the emergence of novel business models.

There are many concepts closely or vaguely related to the DDH – we will classify the most important ones, then describe the requirements for a DDH, focusing on the privacy and security aspects.

We will use the term DDH for a Data Hub that fits our vision, and simply Data Hub for other kind of hubs.

2 TAXONOMY

The DDH vision fits a characteristic place in the landscape of hub-like solutions (Figure 1).

2.1 Non-Hubs

2.1.1 Data Distribution / Data API

This platform allows users (organizations and individuals) to receive data created or assembled by an organization. The external access is basically a read-only API, exemplified by access to mobile movements or financial data gathered by Bloomberg or Refinitiv. This is not considered a Data Hub.

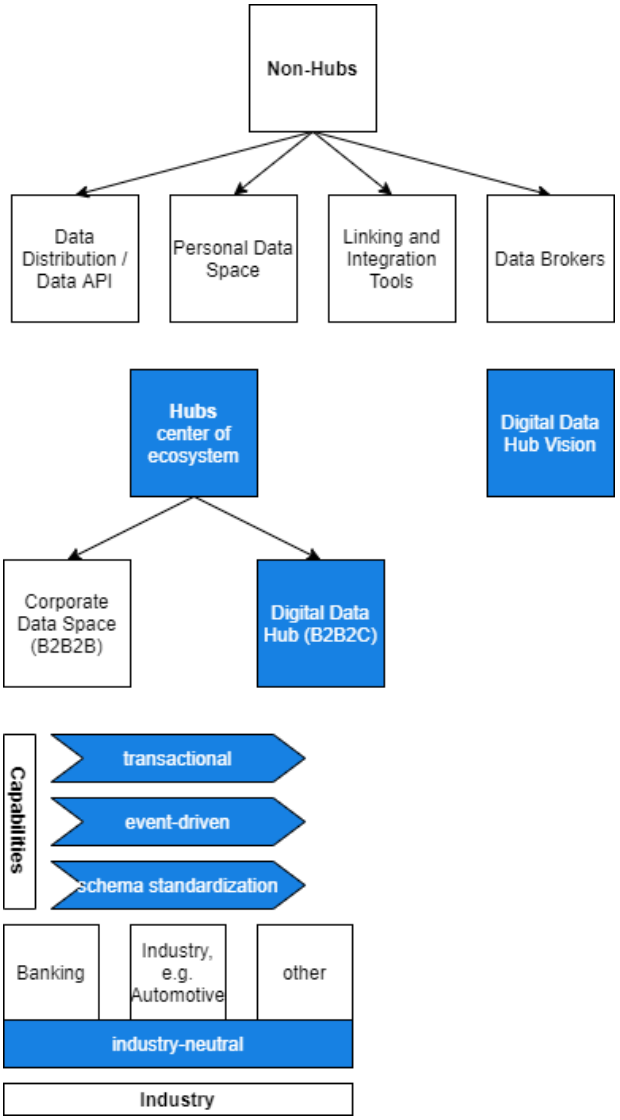

Figure 1: Hub Taxonomy

2.1.2 Personal Data Space

These are spaces for individuals (or smaller organizations, such as small and medium enterprises) to keep individual documents and files, such as document safes. Sharing may be possible, but the ecosystem is limited and there are no transaction facilities.

Examples are Swisscom's now defunct DocSafe, banks' document safes, and the basic offering of Dropbox.

2.1.3 Data Stewards and Brokers

Data stewards and brokers collect data from individual users and market it (or use it for the public good):

- Gener8Ads [34] provides a browser plugin and collects and markets your browsing data, paying back the user and retaining 20%.
- bitsabout.me [22] is a privacy-controlled data collector, where the user can release data sets from each source to selected partners. It intends to profit from the legal requirements for data portability and becomes most useful when data is openly shared by the data originators [17].
- OPAL [14] allows contributed algorithms to access collected data. Algorithms are designed to give answers, instead of giving access to the data itself, thus fostering privacy.

2.1.4 Linking and Integration Tools (Producer-Consumer-Flow)

There are a few convenient tools that link applications and create value on an individual basis (for people or organizations). Examples are commercial offerings such as:

- IFTTT ("If This Then That"): Event-based integrator moving events between sources and destinations [31]
- Zapier: "Zapier moves info between your web apps automatically" [32]
- Integromat: "Integromat lets you connect apps and automate workflows in a few clicks. Move data between apps without effort so you can focus on growing your business." [33]

These systems do provide integrations rather than building an ecosystem.

2.2 Data Hubs

Data Hubs are characterized by ecosystems where multiple parties read and add data for multiple users. Hubs, like their counterparts in transportation, are characterized by an ecosystem of participants, such as main carriers (creating data), users, and services adding value. Typically, there are multiple parties that create data, transform, combine, or analyze it to add value at each step, and finally use it to deliver value.

Most use cases and participants may yet be unknown when the hub is created. New participants may trigger new use cases, and further new participants, leading to an emergent eco system with novel business cases.

Data Hubs have a clear view of ownership of data, with elaborate consent mechanisms. Most hubs will have active components (such as connectors or the Data Apps described in section 5 running within the hub).

2.2.1 Corporate Data Space (B2B2B)

Organizations willing to share data amongst themselves participate in Corporate Data Spaces. The Corporate Data Space often also shares data with outside participants, such as the general public, acting as a Data Distribution Platform.

IDSA [13] is the prime example of a standard in Corporate Data Spaces, Deutsche Telekom's *Data Intelligence Hub* is a prominent offering [26].

2.2.2 Digital Data Hub (DDH)

The DDH links individuals with organizations (Business to Business to Consumer, commonly abbreviated as B2B2C, the hub is the B in the middle). Data can be contributed by all participants connected to the hub and shared with any other participant. All data is attributed to a participant (the data sovereign), and sharing is always by explicit consent. See also *Understanding MyData Operator* [12] for a good introduction on personal hubs and their requirements.

Although not the prime aim, the B2B2C DDH may also be used between organizations (especially if one of them is small, such as an SME), becoming a B2B2X hub; or between individuals, becoming an X2B2X hub. There are very few restrictions in the vision, apart from possibly the questions of cost-bearing.

The motivation for a person to use a DDH may vary; monetizing data may be a factor, but [10] and our own research found the ease of life as more important. The DDH enables *Digital Agents* to act on behalf or in collaboration with the user to help make decisions and do transactions.

An additional motivation for explicit, consented sharing is a reduced lock-in with organizations (typically global giants) that implicitly gather and store data from users. An explicit consent can be withdrawn at any time and granted to another organization, enabling data portability [17].

2.2.3 Additional capabilities

There are some capabilities that either a Corporate Data Space or a Data Hub can have. These capabilities set the DDH apart from data aggregation and monetization systems and make the DDH continually useful in an everyday context:

2.2.3.1 Transactional

A transactional Data Hub allows participants to enter contracts or to invoke actions based on contracts. The action may be initiated by the user, or by the system based on the execution of a smart contract. A transactional Data Hub may be a B2B2B or B2B2C hub.

Transactions are the cornerstone for linking data with the real world in a bidirectional way.

2.2.3.2 Event-driven

An event-driven Data Hub provides data as soon as an event occurs. Contrast this with a dump of past data - which may be driven by rights to your personal data enforced by regulation [17]. Low-latency, real-time control events (e.g., for autonomous driving, or remote surgery) are explicitly excluded and not deemed feasible in a shared Data Hub in the foreseeable future.

2.2.3.3 Scheme Standardization

Most useful data have at least some structure. There are various degrees of standardization of the structures. A Data Hub with some Scheme Standardization capability supports some unification (this is intentionally vague, to avoid excluding too many forms from the start). Section 6 describes the DDH data standardization.

2.2.4 Variants

2.2.5 Identity Hubs / Self-Sovereign Identity

Identity Hubs provide credentials confirming someone's or something's identity. Their implementation may be completely distributed as Self-Sovereign Identity, but the infrastructure may be provided by a hub [4].

In particular, Self-Sovereign Identity platforms may provide provable answers to questions ("zero knowledge proofs"; e.g., answering if a person is over 18 years old) rather than revealing the underlying data (e.g., date of birth) [14]. In contrast, Data Hubs provide full or restricted data sets for explorative analysis. See also sections 7.2 and 7.3.

2.2.5.1 Domain-centric / industry-specific hubs

Open Banking is a typical hub applied to a single domain. See *Open Banking Project* [18] for an overview. In an interview, experts of Finnova (a cooperation partner of Swisscom Banking) say "perhaps the term Open Finance or Open Business would be more appropriate. There are a multitude of use cases from a variety of ecosystems, for example in the areas of mobility, travel or health" [19].

In this sense, the DDH extends *Open Banking* to "*Open Almost Everything".*

2.2.5.2 DDH

The DDH is primarily a B2B2C Data Hub, which is transactional, event-driven, and provides for Scheme Standardization facilities (although not for complete uniform schema standardization).

3 REQUIREMENTS

There are some high-level requirements, which are used to classify existing platforms, as described in section 2 on taxonomy. The high-level requirements should achieve *completeness of vision* (in the sense of the Gartner Magic Quadrant [36]).

It is important to note that these requirements try not to prescribe an architecture, although they impose some constraints (e.g., event-driven API).

3.1 Market and Organisational Requirements

The complete vision of the DDH encompasses the a number of key elements:

An **Organization**

- that uniquely identifies data and assigns it to its owner
- that enables and curates data standards, as described in 6.1 and oversees data sensitivity (7.3)
- that certifies Data Apps (3.3.1)

A **Market**

- for Data Apps and data through them

A **Clearing House**

- for onboarding, contracts and billing
- auditing all consents and accesses

A **Secure Infrastructure**

- as an execution environment for Data Apps
- for their interaction with users

3.2 Openness

If the core platform is open source, it helps to build trust. The Swiss voting systems (Swiss Post, Canton of Geneva) have come under attack because they were not open source; once source code was revealed, weaknesses were found quickly.

3.3 Governance

Trusts or *cooperatives* instead of a commercial enterprise have been proposed as governing body for the DDH [6]. They provide more control and accountability than *commons* would, but are may also have problems of unclear participation rights and requirements, and insufficient funding.

However, the recent security and privacy problems with the Swiss vaccination portal *meineimpfungen.ch* (leading to its sudden shutdown) [35] has shown that a cooperative is not a panacea in this domain. Organization such as Swisscom, the incumbent telecom provider, or the Swiss Post with its majority ownership by the Swiss federation should have the required trust, market base, neutrality, and financial backing.

A monopoly, however, would be problematic; although there are platform effects, the DDH platform must be open to connect to other hubs and be interoperable (see 4).

We expect the industry to move to more decentralized architectures over time, but do not believe that an organization without a central control would be acceptable to Swiss users in the next few years.

3.3.1 Certification

Data Apps must be certified; their source code and external connection requirements must be presented for the certification. The certification process would be similar to the process of popular phone app stores. The certification will not be perfect and the DDH cannot guarantee for their behavior. Data Apps are considered to be semi-trusted; i.e., they need to be run secure environment and closely monitored. The certification may offer different levels of scrutiny, which is declared in the market.

4 FEDERATION AND INTEROPERABILITY

4.1 Federation Requirements

Although the DDH encompasses many organizations and users in many domains, it will not be global and closed offering. The true value of Data Hubs come to the forefront when it enables personal data across Data Hubs to be combined in analytics in a privacy-preserving manner, yielding benefits to the individuals participating. We refer to this as the *federation* of the Data Hubs.

Federating Data Hubs provides an avenue towards supporting the data driven society, which promises greater insight into communities and societies, allowing governments and individuals to improve planning [6]:

- Data has more value when combined across disciplines: Although seemingly self-evident, the weight or impact of this assertion comes to the forefront when we see analytics results (using multi-discipline sourced data) that give us insights into things that were previously unknown or unimagined.

- Data increases in value when it is shared: The WISH report [8] on big data in health identifies data sharing as key to solving the world's health problems (e.g., arising from the spread of diseases due to increased human mobility).

A federation of Data Hubs must address the various technical, legal and policy challenges around personal data stored in the Data Hubs of individuals [6]:

- Access policy administration: Providing the data-owner(s) with semantically meaningful methods to determine access and usage policies for their data
- Access policy enforcement: Implementing the access policies using mechanisms within the Data Hubs that enforce those policies consistently, without ambiguity and without any resulting data loss.
- Access policy tracking and accountability: Providing the data-owners as policy administrators with method and mechanisms to track, record, audit and reconcile data-access events in a semantically correct and consistent manner.

The federation of Data Hubs must permit a user-centric management of access policies to the various datasets individuals own in the Data Hubs.

- *Single Policy Administration Point (PAP):* The participant in the data should be presented with one place in which to configure access policies and grant/revoke authorizations for all the data that she owns. This single PAP approach means that if an individual participates in multiple Data Hubs, then the same policy must be enforced consistent across these separate Data Hubs.
- *Enable distributed logging and audit:* The individual who participates in a Data Hub should be provided with information regarding access (and attempted access) to their data located in the Data Hub. This feature should be enabled for all the Data Hubs in which the individual participates in across the Internet.
- *Facilitate Data Usage Terms*: As part of the Data Hub, the participant in a Data Hub should be provided with tools to enable them to easily select from a selection if standardized data-usage terms (i.e., Terms of Service). The usage terms must comply to the prevailing Data Privacy regulations (e.g., GDPR) and should be crafted with the help of legal experts. This is strictly enforced in the DDH itself; see 7.6.

4.2 Interoperability

Although Data Hubs are far from being standardized at a level comparable to the Internet communications layer (TCP/IP) standards, there are both standard stacks and interoperability mechanisms.

4.2.1 Hub to Hub communications

The open architecture of DDH with its Data Apps allow a seamless integration with other Data Hubs. All outside communications goes through dedicated Data Apps. Data Hubs that are considered to offer equivalent privacy and security protections may use privileged inter-hub Data APIs (see Appendix A).

4.2.2 MyData Operator

The MyData Global [11] organization sets requirements and standards for consumers to access and share data form organizations they interact with. They have a human-centric view of personal data, putting the individual in control of all data exchanges.

A B2B2C Hub should comply the MyData Operator stack as described in [12]. This ensures interoperability with other MyData Operators.

In Switzerland, bitsabout.me [22] and the tourism platform discover.swiss [23] are MyData Operators.

4.2.3 IDS-RAM

The International Data Spaces Association defines a Reference Architecture Model for interoperable data spaces. This is mainly designed for collaboration within an industry or cross-industries, as opposed to consumer connectivity. It emphasizes both data sovereignty, explicit consent, and a certification process for connectors and participants.

For interoperability between B2B hubs, the hub should be compliant with the International Data Spaces Reference Architecture Model [13].

5 DATA APPS FOR PROCESSING DATA

5.1 Data Apps

Data Apps set the DDH apart from simple producer-consumer flows. Basically, every organization can contribute Data Apps. They contribute, transform, and analyze data. They are the primary means to create value from data as well as for data ingress and transmission of transactions.

The DDH must provide a highly constrained, secure environment to run data apps, so they won't have unauthorized access or can publish data derived from data to which they don't have full rights. Yet the environment must be flexible and foster creativity. In particular, Data Apps must be able to work in the *background* (periodically based on a subscription, or event-driven) without the user's presence.

5.1.1 Data App Store

There must be a store-like facility where users can find certified Data Apps and connect to them (see 3.3.1). Data Apps are the primary means to discover data sources and analytics; there is no separate storefront for data.

A Data App may present a HTML user interface, which is framed and monitored as an external connection (see 8.2.5).

5.2 Simplicity and Ecosystem

The requirements of the platform should be simple yet provide for composability among Data Apps. The power of the Data Hub is contributed by the ecosystem, not by the platform itself. Namely, all data contributions, transformations, and insight come from Data Apps.

Good programmers working in any organization must be able to develop Data Apps, using open source environments and tools. No very specialized or advanced knowledge should be required – as this would severely limit the potential participants of the eco system. In particular, Data Apps should be able to work with data in a common form, e.g., JSON.

5.3 API Requirements

All functionalities must be available to Data App developers through simple APIs. There should be no GUI-only functionality. Native web APIs should be used, with some thought given to their evolution. This calls for a standard way to identify each data item, for example, using a URI; for which keys as described in section 6.3 form the base.

5.4 Scalability

The platform should have the potential to scale up to the population of Switzerland, and to all major companies and 50% of the SMEs. Media (video, etc.) are not a target, so YouTube-level storage scaling is not a primary concern.

High-frequency (observing) data moves should be event-driven (this is probably both an API requirement and a consequence of scalability).

6 DATA MODELLING STANDARD

Identifiers describe how a data item is named. Schemas describe how a data item's content is structured.

6.1 Standard Schemas

Well-structured data is essential for seamless processing, but it is also a necessity to understand the data in order to protect it (it is too easy to hide personal data in unstructured data). However, the world of data formats is a wild place. Open Banking is taking off only after standards have been established [19], so they are important.

We envision a **dual-tree scheme**, for data defined by a contributing organization, and data following standards defined by the hub.

6.2 Common core, federated extensions

Federated formats divide a domain into subdomains and establish a standard or a dominant player's scheme for each subdomain. This is often the only feasible approach – but divergent definitions of core concepts are problematic: Subdomains overlap; and non-standardized areas exist.

Common core, federated extensions aim to simplify the federated formats, by covering the essentials of the domain in a common core. Flexibility is provided by extensions to the core, which may overlap. However, extensions must always map to the common core, allowing scheme users to reach into the core, ignoring extensions. Phrased differently, each extension must fully understand the core; whereas the core is not required to know about or understand extensions.

Many successful extensible standards use these principles, such as XML, HTML, TCP/IP.

6.2.1 Organizational Tree

The organizational tree (*/org*) provides for data structured according to an organization's scheme.

It also provides for very loosely structured data, with an identifier and perhaps some tags (schema-less data).

The top levels of the org tree are organizations, which are identified by their root domain (e.g., swisscom.com, migros.ch, admin.ch)

6.2.2 Hub Standard Tree

The hub standard tree *(/p*) provides for data structured according to a standard. The standard is defined by the DDH, with contributions from the ecosystem. The standard schema is not completely covering all domains with the same granularity (that would be an impossible undertaking, even if the data world would be static).

We propose a federated schema, which combines schemas from different domains, usually relying on the market leading schema. In areas where there is no established market leader, the schema may point back into a schema on the Organization tree.

Of course, items on the organization tree may also refer to the Hub Standard Tree in areas where an organization conforms to a standard. It may also extend the common core with additional attributes but must not change the meaning of standardized elements.

For example, a bank deposit in the hub standard tree (under finance) can refer to a specific deposit structure in the organization tree of the bank from which the data is obtained. This avoids standardization that is too early and too specific.

Figure 2 illustrates the two trees *organization org* and *personal p,* with organizational schema entries pointing to standard schema entries, which are based on common standards

6.2.3 Translation between Schemas

Data Apps should be able to translate data between schemes. This is a benefit of the ecosystem – the organization and growing standardization are expected to be provided by interested parties, not necessarily by the hub operator or the data source.

Data leaving the hub should be strictly compliant to a schema, so the risk of hiding personal data is minimized (however, it is still possible to use advanced concealing techniques, i.e., embed the information as an invisible watermark into an image).

6.3 Identifiers

Data Items should be individually addressable by *keys*. In a REST API world, the keys should map to a URI. The key must identify the owner of the data, designated by its principal.

Consent should be grantable on any key level up to the owner identifier.

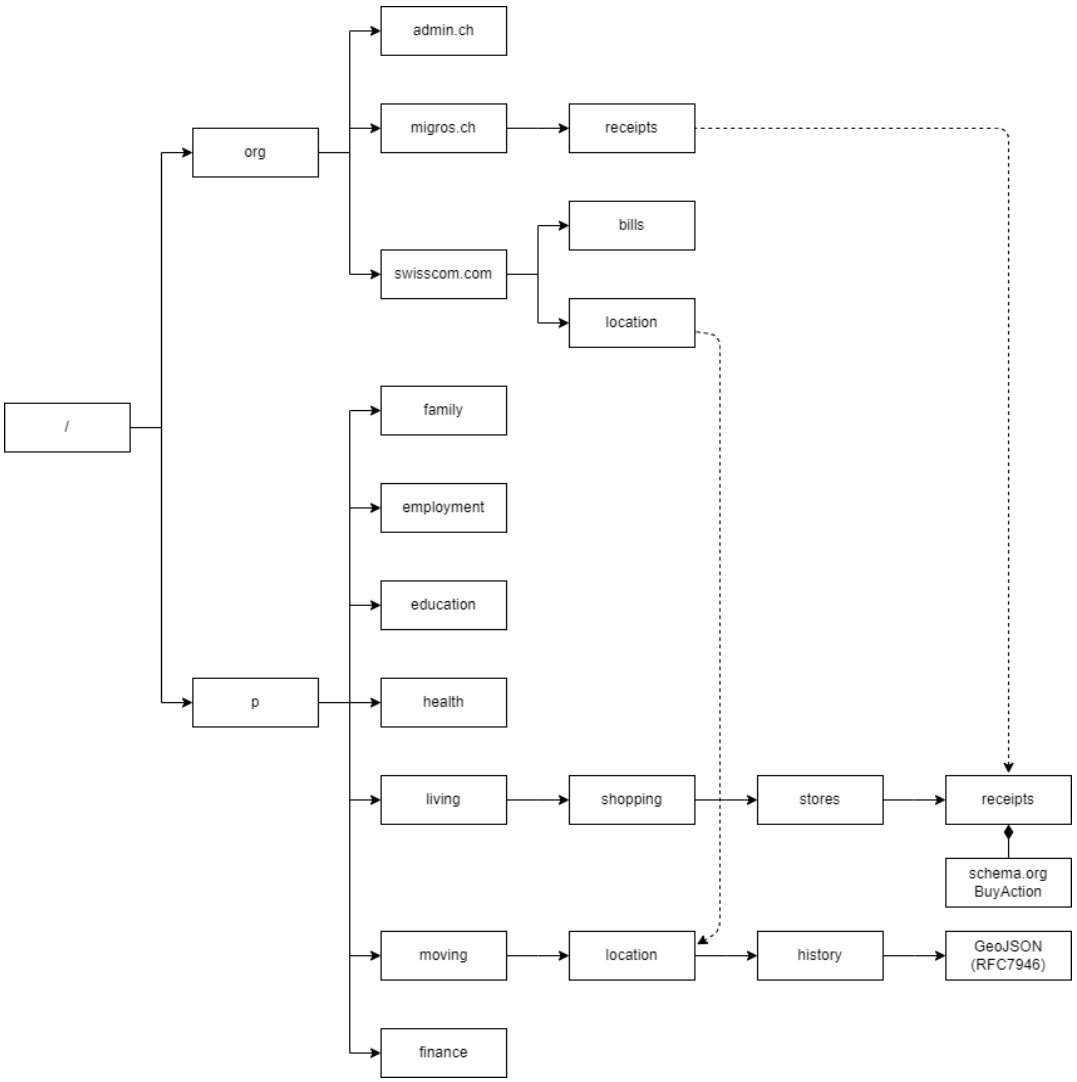

Figure 2: Schema Tree

7 PRIVACY REQUIREMENTS

7.1 Purpose and Privacy by Design

The purpose of the DDH is to allow sharing of personal data with a large number of relations. We will describe the main mechanisms that address the foundational principles of Privacy by Design [1].

A DDH as a collection of data obtained for later sharing may per se contradict data frugality and purpose limitation. The user-centricity and full functionality outweigh this concern: The DDH is an open-purpose system, the purpose of sharing data is often not known in advance.

Privacy as the default setting: All sharing has to be explicit, under the full control by each user, revocable at any time. The partner data is shared with has to be fully identified.

Full functionality: It is important that the DDH controls the security and privacy of data it is trusted with in the same way as a bank controls the security and privacy of money in a bank account, but without preventing the user from working with their data.

7.1.1 Comparison with Big Data Analytics

The DDH may assemble a lot of data, but in a very different way than the Big Data analytics run by many Internet services:

Characteristic	**Big Data Analytics**	**DDH**
Main focus	collective of users	individual person
Source of data	browsing and user actions	transactions with relationships
Purpose	marketing	enable agents to assist user
Consent	implicit by using service	explicit, revokable

7.2 Identity

The identity of the users is one of their most valuable assets. A qualified identity is essential for both sides of a transaction, as a fake identity can lead to loss of assets and enforcements options.

7.2.1 Centralized or federated user identity

If someone wants to share shopping data with the health insurer, both the shop and the health insurer need to establish the person's identity.

As a national electronic identity (E-ID) law has failed in Switzerland [20], a private identity provider must be used. The only identifier covering almost all inhabitants is the social security number (AHV13), but it is not widely used outside of some government relationships. Using it could also expose government data about a user.

Private solutions exist, the most prominent is SwissID, [27]; there is also a widespread solution for verified logins, Mobile ID [28]. OptioPay [25] links the principal to a bank account or a phone number.

7.2.2 Self-sovereign identity

Self-sovereign identity [4] is a natural partner and complement of the DDH, as it may provide identity and user attributes in a privacy-preserving fashion. In particular, Identity Hubs (see 2.2.5) may be an ideal complement to a DDH.

7.3 Data sensitivity

Data attributes can be split into the following designations [2]:

- explicit identifiers of a principal, such as name or social security number
- quasi-identifiers, allowing to re-identify the principal by smart combinations
- sensitive attributes (such as salary, diseases)
- non-sensitive attributes, i.e., all other attributes

Data attributes can only be categorized if there is a well-understood schema associated with it. All attributes that potentially reveal identity must be declared in the schema.

When data is shared anonymously or pseudonymously (see 7.4 below), these attributes must be encoded or removed before data is transferred.

Certain revealing fields of the schema (e.g., date of birth) are converted into less revealing fields, e.g., age group or age flags (e.g., >18y), by a process known as *Schema Generalization.* Each transform should result in a weaker designation.

While identifying attributes must be excluded or transformed for anonymous sharing, attributes or whole subtrees under an identifier (see 6.3) may be designated as sensitive, and require explicit consent to include them in sharing.

Attribute and data classification is a major task for the DDH operators (see 3.1); it must be simple but have an input and review process.

7.4 Anonymization and Pseudonymization

Respect for user privacy: While contributing *anonymized* data has its value, it is not possible for agents to provide results based on anonymous user data.

Pseudonymization, where the hub knows the identity of the user, but doesn't rely it to a Data App, allows for specific offers, answers, and recommendations by agents. The identity might be revealed later (by explicit consent) but is not revealed in an offer exploration scenario.

This is time-tested mechanism on the form of box number advertisement in newspapers and magazines.

Anonymization and Pseudonymization without accidentally revealing the identity is an art and a science.

Concepts such as differential privacy [2] (for the anonymous contribution of data for statistics collection) and confidential computing [1] (for analysis without

revealing of identity) may be used. However, the environment for Data Apps must still be simple and versatile enough so that average developers in startups can use it.

7.5 Consent

Consents are explicitly granted by the principal owner of the data to another principal, which is an organization, or another user. Consents carry restrictions about what can be done with the data (e.g., use to combine with data from others; share pseudonymously), and what the highest sensitivity of data to be shared may be.

7.5.1 Dynamic Consent

Consent may not only be given to a principal, but to a group of principals sharing common criteria. For example, to obtain offers from insurers, consent may be given to all insurers, provided the list of insurers is made transparent and frozen at the time consent is given.

7.6 Terms and Conditions

Visibility and transparency: There should be just a small number of easy-to-understand Standard Terms and Conditions, linked to privacy icons [21]. The Terms and Conditions shall not be customizable, so users don't have to spot differences in the fine print.

7.7 Data Offers

Data Offers to a principal owner of data consist of the data required (stated as consents required), the Terms and Conditions, and the price being asked or offered.

7.8 Contracts

Once the principal owner of the data accepts an offer, it becomes a contract, and the consents are being granted.

Subject to its Terms and Conditions, the Contract can be terminated at any time. Terms and Conditions which involve a delay until a cancellation gets effective are discouraged and must be reserved for special-purpose deals.

7.9 Negotiation and Brokerage

If several Data Apps are orchestrated in a process, the DDH may offer a Brokerage Service (as a Data App itself) to negotiate acceptable Offers between all parties involved.

7.10 Joint ownership of data

Banks have a long tradition and established practices for accounts with joint access. Either each principal holder can initiate transactions, or all holders or a quorum of the holders are needed.

The same mechanism applies to Data Accounts; consent can be given by one of the holders, by a subset or quorum, or by all holders. A Data App implementing workflows may be provided to assist the users in the administration.

7.11 Privacy Evaluation

We evaluate the privacy solution against the principles of Privacy by Design [1]:

Principle	**Solution**
Proactive not reactive; preventive, not remedial	Data is intentionally provided, not coincidentally collected.
Privacy as the default setting	All data is private unless access is granted by explicit user action.
Privacy embedded into design	Secure sharing of known data is at the core of the design.
Full functionality – positive-sum, not zero-sum	Users have all means to use their data and enter transactions.
End-to-end security – full lifecycle protection	See section 8.
Visibility and transparency – keep it open	Transparent, standardized contracts without hidden claims.
Respect for user privacy – keep it user-centric	Pseudonymization provides for full-functional, user-centric privacy.

8 SECURITY REQUIREMENTS

8.1 Isolation of Processing

As the user gives consent to organizations to use data, it becomes the responsibility of these organizations to use the data according to the consent. Bitsabout.me declares this to be the responsibility of the organization with whom the user consents to share the data. Pryv [24] delegates cross-account aggregation and consolidation to "middleware". By delegating the responsibility, an ecosystem of Data Apps becomes as weak as its weakest Data App or its provider. Therefore, Data Apps should be forced to abide by the privacy by design rules.

There are basically two models for running Data Apps:

1. Data Apps run anywhere, and access hub data through the API. Once they obtain data, the DDH no longer has control over it.
2. Data Apps run in a *"walled garden",* whose entry and exits (ports) can be restricted and monitored. Privacy can be compromised by side doors or improper combination and attribution of data.

Only the second approach has a chance to control access as long as the data doesn't leave the "walled garden". If a Data App leaks data, the trust in the DDH itself may be impacted or lost.

Therefore, Data Apps need to be running in an environment controlled by the DDH, so their data access can be controlled and their interaction with the outside world can be restricted and monitored.

8.1.1 *Edge (on user device) processing*

We consider data storage and processing on the user device as interesting from a security point of view, but not practical for the many relationships the DDH envisages. However, some user may choose to keep their private keys (see 8.3) on their own device.

8.2 *Walled Gardens*

All processing is isolated in an operating system parcel (e.g., a Container), access to hub data is only available through-well defined APIs, and outside access is restricted. The resulting isolation is at least at par with current banking systems (Figure 3).

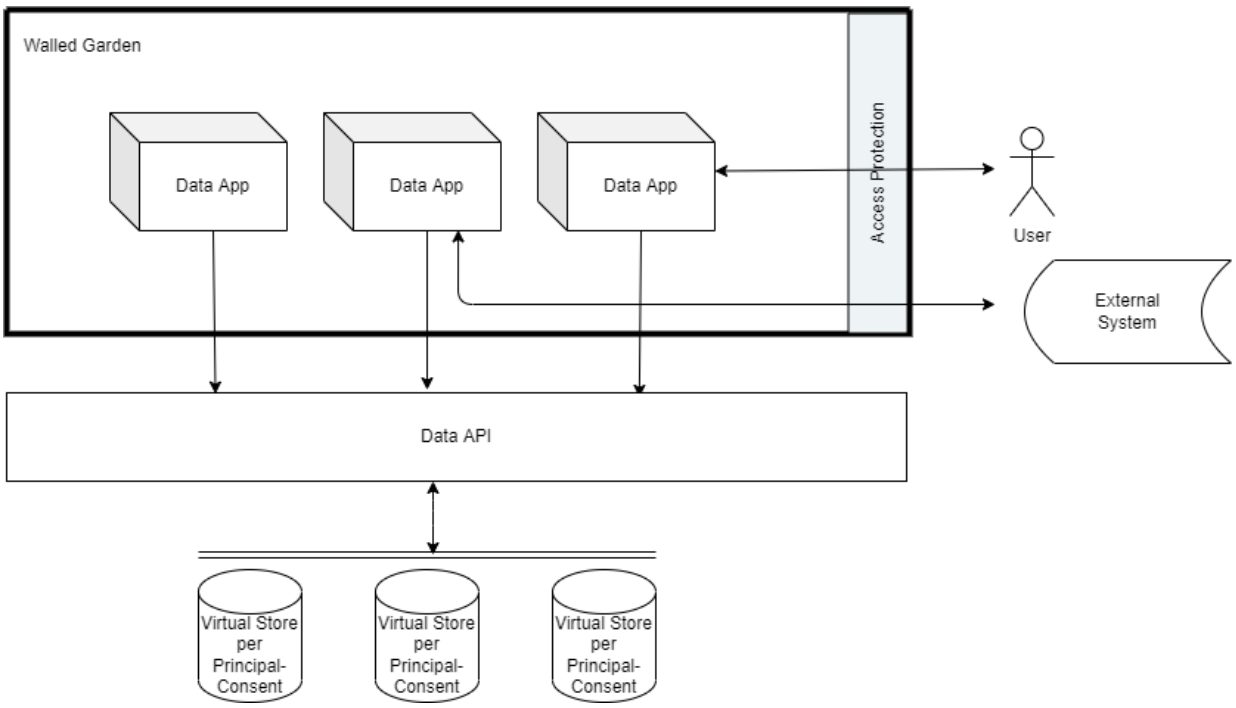

Figure 3: Walled Garden

This is the solution we will discuss further.

8.2.1 *Data Access APIs*

Data Apps interact with the DDH through Data Access APIs. For each interaction, the Data App must declare the involved user, and its intent of the data access: Reading, writing, anonymous or pseudonymous access, etc. Each Data Access is checked against the consents, and audited. See Appendix A for the list of API operations.

8.2.2 *Auditing*

Keeping an audit log of each access and each change of consent is important to establish the correctness of operation and to track violations. The audit log must be safe from later changes and must not be repudiable; digital ledger technologies are best suited for this.

8.2.3 *Secure computing and enclaves*

The most secure environments (apart from air-gapped facilities) are confidential computing environments or enclaves [29]. Such technologies help to build walled gardens (see 8.2) and to protect against side-channel attacks (see 8.2.4).

They may not provide a complete execution environment to host Data Apps contributed by third parties without unduly restricting their programming paradigm. However, they may host the most critical parts, especially the controller of the Walled Garden.

8.2.4 *Protection against side-channel memories*

Malicious Data Apps may retrieve data for one user, retain it in its memory, and use it later on for the benefit of another user. This means that state is shared between users. This is a major issue when semi-trusted Data Apps are run on the DDH.

Clearing memory and temporary files between transactions involving different users is possible using Confidential Containers [29] and a careful management of memory, temporary files and execution state by the Walled Garden controller.

The Data App (running in a container) knows at which point a new request can be handled; assuming it is a REST service, it is the point accepting calls. It then requests the Walled Garden to save its memory. The Walled Garden does this, provided no data requests have been made beforehand by the Data App (so there is no data lying around).

The Data App can itself request for reinitialization at any time, at which point the Walled Garden replaces its memory, temporary files and execution state, and resumes its execution. The Data App can ask the Walled Garden to replay an idempotent request after the reinitialization; this may be used to replay the request which caused the reinitialization.

If Data App asks the Walled Garden for conflicting data (data from two principals which must not be retained), the Walled Garden instead tells the Data App to request a mandatory reinitialization.

The diagram in Figure 4 illustrates the state being saved (1), a request being handled (2), another request (3, in green) causing a reinitialization (4, in blue), with a replay of the request (5), which then obtains data and proceeds:

The Walled Garden should aim to schedule requests so that the number of mandatory reinitializations is minimized. Basically, it should schedule requests for the same principal on the same containers (affinity), with no intervening requests for other principals.

8.2.5 External Communications

External communication, including the user interaction, must be restricted to designated Data Apps running in their own containers. The Walled Garden acts as a proxy for them. Standard firewall and data-leak-prevention techniques must be applied to these connections. A boundary control tool such as Cyral [30] may be used to protect ingress and egress.

8.2.6 Caches

All data flowing between Data Apps is cached in a special section of the tree. The Consents are always inherited from the original data, and cannot be modified.

There may be *events* carried directly between Data Apps, but they may not carry any data except keys into the tree.

8.3 Protection of data at rest and on the move

All data at rest (in databases and caches) and in transit in and out of the hub must be protected from being accessed either by the hub itself or by any third party that was not given consented access to the data.

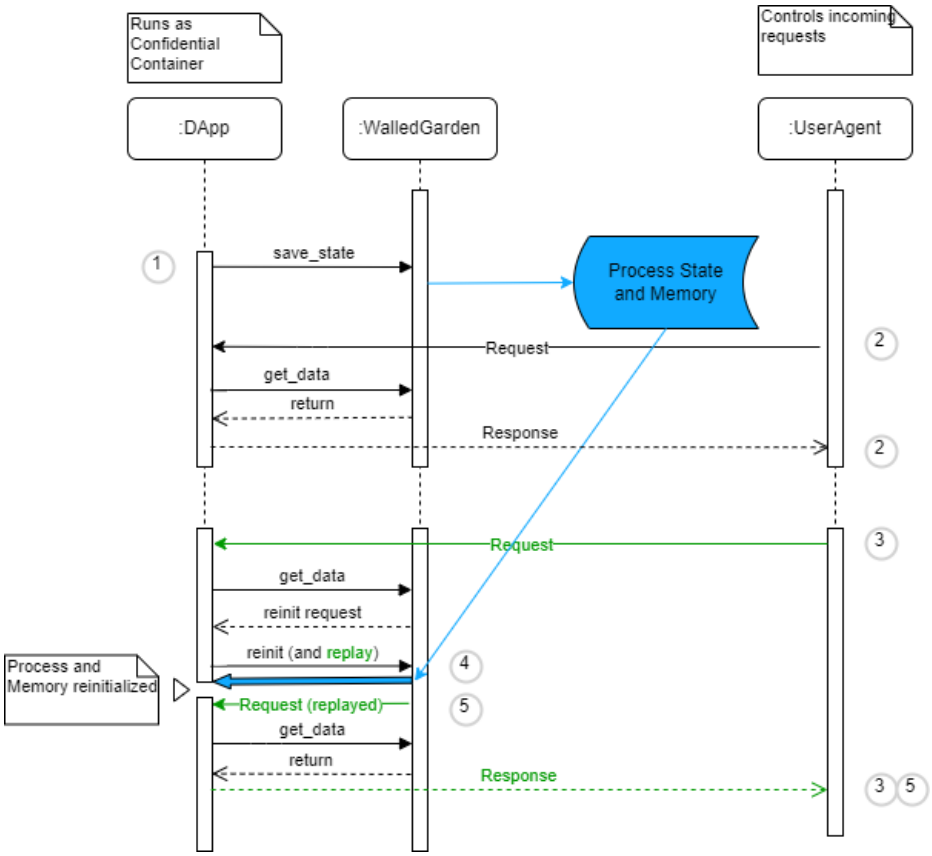

Figure 4: Processing Data for different principals

Bitsabout.me and Pryv achieve this by encrypting the data account of each user with a key that is in the sole possession of the user (and stored on the user's device). While this prevents a massive theft of all user data, it also inhibits background processes (like agents) from working on behalf the user. It delegates most security aspects to the sources and targets of the data. OptioPay [25] keeps a central key for all data in secure storage.

The DIF Secure Data Storage Working Group [15] is working on standards to provide encrypted data vaults.

For the DDH, we note that the data must be accessible by their owner and all parties the owner has given consented access. The DDH itself does not need and indeed should not be able to decrypt the data. Incoming data events concerning a user can be forwarded to the Data Apps based on the consents alone.

There are always at least two parties involved in transmission of data and in its storage, namely the owning principal (which we call Alice or A) and the principals given consent (the first of which we call Bob or B).

Initially, the tree T owned by Alice is encrypted $\{T\}_N$ using a random nonce N, an ephemeral key that is generated by the DDH, but never stored. The nonce is then encrypted using Alice's public key and stored in the DDH as $K_A+\{N\}_{K_A}$, along with Alice's public key.

When Alice gives consent to Bob for a certain part of the tree, all data at or below the point where she wants to give consent will be re-encrypted with a new nonce $\{N'\}$. That new nonce which will be encrypted and stored twice, once with Alice's public key $K_A+\{N'\}_{K_A}$ (so she retains access) and once with Bob's public key $K_B+\{N'\}_{K_B}$ (so he gains access).

For the re-encryption, Alice is asked to decrypt the nonce with her private key, and hence decrypt the data. Whether Alice keeps her private key on a device or entrusts the DDH is not relevant to this discussion. She needs to use her private key when she gives consent; Bob's private key is not required at that time (as his public key is known to the DDH). Bob can use the consented data with his private key without Alice being present.

When Alice gives consent for the same data to additional principals, or if she withdraws the consent for Bob, the same re-encryption applies. Bob loses access when his encrypted copy of the nonce is no longer able to decrypt the data.

8.4 Security evaluation

Security is a multi-faceted endeavor, and complex solutions may make matters worse [3]. Massive thefts of private data would destroy the reputation of the DDH immediately (they correspond to the *maximum credible accident* in nuclear energy, but seem to occur more frequently), so we need to combine privacy, security and a good governance model.

The combination of the techniques outlined here provide enough security for most data, and exceed current standards used in many sensitive domains. The security solutions need not to be on the counterespionage level, but proven and sound.

The weakest spot is usually the system boundary, which must be isolated and monitored. See Appendix B for a list of attack vectors and proposed mitigations.

9 INTERNATIONAL ASPECTS

This paper focuses on data hubs for Switzerland. However, the need for data is not only crucial for the running of the local economy, but also for the global economy. For example, much of international trade involves the exchange of data among the multiple stakeholders in the ecosystem (e.g., exporter/importer of goods; national governments; maritime shippers, etc.).

Decisions regarding investments by international entities (e.g., multi-national corporations) in local industries often require insights into the industries and the communities around them. On many instances, this requires insights to be computed based on data held by public sector organizations (e.g., Bureau of Statistics) and by private sector organizations (including data held by individuals in the Data Hub).

The *sharing of insights* (versus the sharing of raw data) provides a way to provide the relevant insights to parties who need these insights for decision-making, without giving them direct access to the underlying data [6].

In the case of international parties (e.g., entities located in a different legal jurisdictions), the controlled sharing of insights is required while satisfying a number of general privacy-related requirements:

- *Clear identification of the source of Data Apps and algorithms*: Data Apps, including Algorithms, which have been submitted to a Data Hub (for execution there), must include unambiguous identification regarding their source. For example, Data Apps (e.g., Python code) should be digitally signed, with an accompanying digital certificate that identifies the source.
- *Proxying by local (domestic) entities*: In the case of international parties, a domestic entity agreed upon by the owners of the Data Hub could act as a proxy, receiving the Data Apps, verifying them and then passing them to the Data Hub. See also section 4 on Federation.
- Protection of responses: Insights pertaining to groups of citizens and domestic communities may carry geo-political significance. As such, a secure delivery method must be employed when returning responses to international parties.

- Express consent from the citizen: Some citizens may feel uncomfortable having their personal data participating within in aggregate computations whose insights are intended for foreign recipients. As such, clear and express consent may need to be obtained when data hubs are employed in this manner.

10 CONCLUSION

We have shown that a DDH with full functionality gains its power through Data Apps, which can be contributed by organization of any size. Flexible standardization is the key to privacy and security in such an open and democratic structure, which aims to be the digital operating system connecting individuals and organizations.

11 ACKNOWLEDGMENTS

Martin's thanks go to the colleagues at Swisscom and in the data community for their encouragement, and especially to Christian Schüpbach for sponsoring this work and for valuable feedback. We also thank Professor Omar Al-Kadi (University of Jordan) for his feedback.

12 REFERENCES

[1] W. Stallings: *Information Privacy Engineering And Privacy By Design,* Addison-Wesley Professional, 2019.

[2] B.C.M. Fung et al.: *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*, Chapman & Hall/CRC, 2011.

[3] R. Anderson: *Security Engineering: A Guide to Building Dependable Distributed Systems,* Wiley, 2021.

[4] A. Preukschat, D. Reed: *Self-Sovereign Identity,* Manning, 2021.

[5] T. Hardjono, D.L. Shrier, and A. Pentland: *Trusted Data,* MIT Press, 2019.

[6] A. Pentland, T. Hardjono: "2. Data Cooperatives" in *Building the New Economy*. https://doi.org/10.21428/ba67f642.0499afe0 (accessed Dec. 3, 2021).

[7] T. Hardjono, "Owner Centric Access Management for IoT Data: Incentivizing Data Owners to Share Data into the Data Markets" in *New Solutions for Cybersecurity* (ed. H. Shrobe, D. Shrier and A. Pentland), MIT Press 2017, pp. 405-422.

[8] A. Pentland, T. Reid, and T. Heibeck, *Big data and Health – Revolutionizing medicine and Public Health: Report of the Big Data and Health Working Group 2013*, World Innovation Summit for Health, Qatar Foundation, Doha, Qatar, 2013. https://kit.mit.edu/sites/default/files/documents/WISH_BigData_Report.pdf (accessed Nov. 27, 2021).

[9] A. Pentland, *Saving Big Data from Itself,* Scientific American, No. 311, pp.64-67. August 2014.

[10] Vodafone Institute for Society and Communications: *Big Data – A European Survey on the Opportunities and Risks of Data Analytics*, Berlin, 2016. https://www.vodafone-institut.de/bigdata/links/VodafoneInstitute-Survey-BigData-Highlights-en.pdf (accessed Nov. 27, 2021).

[11] MyData Global ry: *My Data,* MyData Global ry, Helsinki, Finland. https://mydata.org (accessed Nov. 27, 2021).

[12] J. Langford, A. Poikola, W. Janssen, V. Lähteenoja, M. Rikken (Eds.) *Understanding MyData Operators*, MyData Global ry, Helsinki, Finland, 2020. https://mydata.org/wp-content/uploads/sites/5/2020/04/Understanding-Mydata-Operators-pages.pdf (accessed Nov. 27, 2021).

[13] B. Otto et al.: *International Data Spaces Reference Architecture Model*, Version 3.0, International Data Spaces Organization, Berlin, April 2019. https://internationaldataspaces.org/download/16630/ (accessed Nov. 27, 2021).

[14] Opal Project: *Open Algorithms,* Boston MA. https://www.opalproject.org/about-opal, (accessed Nov. 27, 2021).

[15] Decentralized Identity Foundation: *Identity Foundation: Secure Data Storage Working Group*, Decentralized Identity Foundation, San Francisco. https://identity.foundation/working-groups/secure-data-storage.html (accessed Nov. 27, 2021).

[16] Network Digital Self-Determination: *Discussion paper on digital self-determination*, Berne, Switzerland, June 2020. https://digitale-selbstbestimmung.swiss/home/en/245-2/ (accessed Dec. 4, 2021).

[17] A. Golliez, C. Laux et al.: *Guideline for the implementation of data portability in Switzerland*, Swiss Data Alliance, Zurich, Switzerland, August 2020. https://www.swissdataalliance.ch/publikationen-content/2020/8/26/leitlinie-zur-umsetzung-der-datenbertragbarkeit-in-der-schweit (accessed Nov. 27, 2021).

[18] Open Banking Project: *Open Banking Project*, St. Gallen, Switzerland. https://www.openbankingproject.ch/en/ (accessed Dec. 3, 2021).

[19] S. Biellmann, R. Hutter: *Open Banking: Ein Jahr nach PSD2 – wo steht die Schweiz?* Finnova Inc, Lenzburg, Switzerland, Sep 2020. [online video] available: https://www.youtube.com/watch?v=N3NV67z90ts (accessed Nov. 30, 2021).

[20] Swiss Confederation, Department of Justice and Police: *Elektronische Identität: das E-ID-Gesetz*, Berne, Switzerland, March 2021. https://www.ejpd.admin.ch/bgeid (accessed Nov. 27, 2021).

[21] J. Hotz, M. Glatthaar: *Privacy Icons*. https://privacy-icons.ch/en/ (accessed Nov. 27, 2021).

[22] C. Kunz: *Bitsabout.me*, Berne, Switzerland. https://bitsabout.me/en/about/ (accessed Nov. 27, 2021) and private communication.

[23] Genossenschaft discover.swiss: *discover.swiss*, Zurich, Switzerland. https://discover.swiss/#plattform (accessed Nov. 27, 2021).

[24] Pryv SA: *Pryv,* Lausanne, Switzerland. https://www.pryv.com/about-pryv/ (accessed Nov. 27, 2021).

[25] OptioPay GmbH: *OptioPay,* Berlin, Germany. https://www.optiopay.com/ (accessed Nov. 27, 2021); and private communication.

[26] Deutsche Telekom IoT GmbH: *Data Intelligence Hub,* Bonn, Germany. https://dih.telekom.net/en/ (accessed Nov. 27, 2021).

[27] SwissSign Group AG: *SwissID*, Glattbrugg, Switzerland. https://www.swissid.ch/ (accessed Nov. 27, 2021).

[28] Swisscom (Switzerland) Ltd: *Mobile ID*, Ittigen, Switerland. https://www.mobileid.ch/en (accessed Nov. 27, 2021).

[29] M. Russinovich et al.: *Toward Confidential Cloud Computing,* CACM, June 2021.

[30] Cyral Inc: *Cyral,* Redwood City CA. https://cyral.com/ (accessed Nov. 27, 2021).

[31] IFTTT Inc.: *IFTTT*, San Francisco. https://ifttt.com/developers, (accessed Nov. 27, 2021).

[32] Zapier Inc.: *Zapier*, San Francisco. https://zapier.com/ (accessed Nov. 27, 2021).

[33] Integromat s.r.o.: *Integromat*. Integromat s.r.o., Prague. https://www.integromat.com/ (accessed Nov. 27, 2021).

[34] Gener8 Ads Ltd.: *Gener8Ads*, Gener8 Ads Ltd. London. https://gener8ads.com/ (accessed Nov. 27, 2021).

[35] A. Fichter, P. Seemann: *Wollen Sie wissen, womit Viola Amherd geimpft ist?* ("do you want know what Viola Amherd [a Swiss minster] was vaccinated with?"). Die Republik, Zurich, Switzerland, March 2021. https://www.republik.ch/2021/03/23/wollen-sie-wissen-womit-viola-amherd-geimpft-ist (accessed Nov. 27, 2021).

[36] C. Drew et al.: *How Markets and Vendors Are Evaluated in Gartner Magic Quadrants*, Gartner Inc., Stamford, CT, 2019. https://www.gartner.com/en/documents/3956304/how-markets-and-vendors-are-evaluated-in-gartner-magic-q (accessed Dec. 2, 2021)

Appendix A: DATA API OPERATIONS

Notation

D	data item
O	owner, a principal
C	consent, or bag of consents, always associated with an owner
D[O]	data item owned by O
D[O,C]	data item owned by O, with consent C
E	encrypted data
A	anonymized data
P_p	pseudonymized data, p representing the owner in the data
K	storage key

Operations

Category	Operation	Signature	Comments
Transit (ingress/egress)			Event-driven operations: Subscriptions followed by obtain or push into hub are basically *obtain* operations. Event-driven publish is equivalent to *publish*.
	Obtain from source	obtain(dapp,O) → D[O]	All external systems are accessed by their associated Data App
	Publish to target	publish(D[O], dapp)	
	Transfer to another hub	transfer_to(D[O], hub : DApp)	Owners and consents are preserved and mapped. There is a Data App implementing some "inter hub protocol".
	Obtain from another hub	transfer_from(hub : DApp) → D[O]	Owners and consents are preserved and mapped
Data at Rest			
	Store into permanent store	store(D[O]) → K	
	Load from permanent store	load(K) → D[O]	
	Cache	to_cache(D[O]) → K	
	Obtain from Cache	from_cache(K) → D[O]	
Ownership and Consent			
	Add Owner	add_owner(D[O_1],O_2) → D[O_1,O_2]	
	Remove Owner	remove_owner(D[O_1,2], O_2) → D[O_1]	This is dangerous and not normally done. At least one owner must remain.
	Add Consent	add_consent(D[O,C_1],C_2) → D[O,C_1+C_2]	
	Remove Consent	remove_consent(D[O,C_1+C_2],2) → D[O,C_1]	
Process			
	Transform plain	transform(f,D[O]) → D[O]	
	Transform under decryption	transform_encrypted(f,E[O]) → E[O]) == decrypt(E[O]) | transform(f,D[O]) | encrypt(D[O]) → E[O]	

Category	Operation	Signature	Comments
	Transform encrypted	transform_confidential(fcc,E[O]) → E[O]	processing encrypted data, Confidential Computing
	Split item	split(D[O]) → D_1[O], D_2[O]	
	Combine items of the same owner	combine(D_1[O], D_2[O]) → D[O]	
	Combine items of several owners	combine(D_1[O_1], D_2[O_2]) → D[O_1,O_2]	Leads to join ownership
Anonymization and Pseudonymization			
	Remove owner-identifying attributes, ownership itself remains	anon(D[O]) → A[O]	
	Anonymize collection	anon_collection(D_1[O_1], D_2[O_2],...] → [A_1[O_1], A_2[O_2],...]	Collections may have advanced anonymization techniques such as Differential Privacy.
	Substitute owner-identifying attributes, retain map	pseudo(D[O]) → P_p[O], p→O	Pseudonym identifier *p* is specific per connector and an interaction identifier (extended session, to be studied).
	Identify pseudonymized (un-substitute)	identify(P_p[O], p→O) → D[O]	

Appendix B: ATTACK VECTORS AND MITIGATION

Operation Category / Attack place	Operation	Attack	Mitigation
Transit: ingress	obtain		
		fake source / man in middle	TLS, certificates
		access data of the wrong user	source responsibility
		access unauthorized data	source responsibility
		access too much data	Heuristic monitoring and throttles
	transfer_from	problems in other Data Hub	assume certified Data Hubs are trustworthy
Rest: store	store, load, to_cache, from_cache		
		steal data	Encrypt data at rest
		steal keys	Keep keys in banking-standard HSM.
Process			
Process: inside attacks (by Data Apps)			
		manipulate consents	consent check outside of walled garden
		ignored consent in combining data	consent check outside of walled garden
		out of band combining data (retain data in process)	Process / container reinitialization (8.2.4)
	anon	incomplete anonymization	Schema truncation, generalization
	anon_collection	incomplete anonymization	Differential Privacy for a data collection
	pseudo	incomplete pseudonymization	Schema truncation, generalization
Process: attacks from outside			
		retrieve from process memory	Enclaves / Confidential Computing for very sensitive data.
		write to process memory or code	Enclaves / Confidential Computing for very sensitive data.
Transit: egress	publish, transferer_to		
		fake destination / man in middle	TLS, certificates
		hide data in other data and other schema manipulation	strict schema validation, signed content