Bus Inf Syst Eng 65(5):577–595 (2023)
https://doi.org/10.1007/s12599-023-00816-9

RESEARCH PAPER

A Reference System Architecture with Data Sovereignty for Human-Centric Data Ecosystems

Simon Scheider · Florian Lauf · Frederik Möller · Boris Otto

Received: 14 May 2022 / Accepted: 13 April 2023 / Published online: 22 May 2023

Abstract Since the European information economy faces insufficient access to and joint utilization of data, data ecosystems increasingly emerge as economical solutions in B2B environments. Contrarily, in B2C ambits, concepts for sharing and monetizing personal data have not yet prevailed, impeding growth and innovation. Their major pitfall is European data protection law that merely ascribes human data subjects a need for data privacy while widely neglecting their economic participatory claims to data. The study reports on a design science research (DSR) approach addressing this gap and proposes an abstract reference system architecture for an ecosystem centered on humans with personal data. In this DSR approach, multiple methods are embedded to iteratively build and evaluate the artifact, i.e., structured literature reviews, design recovery, prototyping, and expert interviews. Managerial contributions embody novel design knowledge about the conceptual development of human-centric B2C data ecosystems, considering their legal, ethical, economic, and technical constraints.

Accepted after three revisions by the editors of the special issue.

S. Scheider (✉) · F. Lauf · F. Möller · B. Otto
Fraunhofer Institute for Software and Systems Engineering, Speicherstraße 6, 44147 Dortmund, Germany
e-mail: simon.scheider@isst.fraunhofer.de

S. Scheider · B. Otto
Chair for Industrial Information Management, Technical University Dortmund, Joseph-Von-Fraunhofer Straße 2-4, 44227 Dortmund, Germany

F. Möller
Chair for Information Systems and Data-Driven Enterprise, TU Braunschweig, Rebenring 58a, 38106 Brunswick, Germany

Keywords Data ecosystems · Data sovereignty · Design science research · Personal data markets · Reference system architecture

1 Introduction

The rapid digitization of large parts of human life is based on the proliferation of technologies producing and processing masses of personal data (PD) (Leidner and Tona 2021). By systematically utilizing PD, those consistently advancing technologies [e.g., Big Data and methods of profiling, scoring, and tracking (Birch et al. 2021)] constitute the basis for many data-driven business models of the information economy (Oehler 2016). This entails questions of designing and organizing them in a way that humans can take full advantage (Spiekermann 2016). However, the applied technologies commonly refrain from integrating humans. Rather, they process PD generating significant profits without individuals being aware of this, let alone having any control functionalities. The fact that the increasing dissemination and industrialization of such technologies is detrimental for humans has resulted in the adoption of the General Data Protection Regulation (GDPR) in Europe in order to provide both a state-of-the-art and restrictive legal framework for the advancing digitalization of the rapidly evolving data economy (Aseri 2020; Metzger 2020). However, albeit intended otherwise, the European attempt to protect human rights to PD decidedly lacks an innovative perspective. This impedes the creation of a uniform economic and legal framework throughout Europe that enables joint and sovereign utilization of humans' PD (Oehler 2016). Thus, an adequate amendment, supplementary to data protection, is required that encompasses specified usage permissions for the

(commercial) processing of anonymized and non-anonymized PD, together with the associated exploitation rights (Metzger 2020). In this context, attention must be paid to the practical implementation of data processing concepts of organizations while ensuring fair participation of human data subjects regarding any profit resulting from the authorized processing of their data. Moreover, the required framework should promote data portability and interoperability. Yet, such recommendations, as proposed by several consumer commissions in Europe, have hardly penetrated the consciousness of (European) policymakers (Oehler 2016). This is caused by European data law whose creators have considered the data economy, i.e., data processing organizations, as the profiteer of data utilization only. In contrast, humans are merely assumed to have an interest in data protection and privacy (Oehler 2016). Consequently, the GDPR entirely ignores the economic interests of European citizens and impedes both the systematic generation and the fair distribution of economic exploitation potential related to their PD (Lauf et al. 2022). As a result, there are hardly any developments toward (eco-) systems for the joint monetarization and utilization of PD in B2C contexts (Koskinen et al. 2019), and the GDPR has predominantly failed so far to achieve its objective of facilitating data-driven business models and innovation. To that end, research is urgently required to investigate and design alternative concepts that can be adapted to the deficient legal framework and which enable a joint utilization of PD based on the GDPR. Concurrently, they must ensure humans their rights to personal data and a fair share of economic profits. We consider a reference system architecture (RSA) as the best-possible instantiation of such (missing) concepts, as these models represent ideal–typical solutions for a class of architectures (Cloutier et al. 2009) and thus are appropriate to close this research gap. Accordingly, we narrow down our first research question (RQ) as follows:

- **RQ1**: *What is an abstract RSA for a human-centric B2C data ecosystem that is technically feasible, complies with European data law, and is usable by individuals?*

Our RSA addresses the identified research problem by generalizing components, functionalities, concepts, and processes required to realize data ecosystems centered on humans and their PD. To this end, it resides in a higher abstraction level than common architecture models of concrete IS solutions. This enables the RSA to adequately solve the research problem as it provides an overview of high-level architecture specifications defining crucial design elements of human-centric B2C data ecosystems. In our study, we define *design elements* as all objects, components, or principles used to describe the RSA from an abstract design perspective (e.g., roles, modules / functions, system components). Following Möller et al. (2020), we use the RSA to aggregate generated architectural design knowledge with generally valid design principles (DP). As a result, our second research question reads as follows:

- **RQ2**: *What generally valid DPs for human-centric B2C data ecosystems can be inferred from the RSA?*

In general, our study is of merit for various topics ranging from privacy and trust to the seemingly contrary field of personal data markets, including considerations about human control (i.e., data sovereignty) and their integration into novel information system architectures (i.e., data ecosystems). In particular, we contribute to data ecosystem research from both an academic and a practical perspective. We provide detailed design knowledge which is accumulated in our artifact in order to transfer this evolving concept to a B2C context. In Sect. 2, the article proceeds with the *theoretical foundations* for our artifact, covering fundamentals of (B2C) data ecosystems, the associated concepts of data sovereignty and personal data markets, as well as related work. In Sect. 3, we outline our *research design*. In Sect. 4, we present our *reference system architecture* for human-centric B2C data ecosystems proposing an answer to RQ1. Section 5 comprises an *artifact demonstration* describing its functionality based on a prototypical instantiation. In Sect. 6, we present our qualitative *artifact evaluation* and its results. In Sect. 7, we reflect on what we have learned and *formalize design knowledge* accumulated in the design process to answer RQ2. Section 8 closes with a discussion of the final artifact, elucidating its main contributions, outlining study limitations, and proposing recommendations for future research.

2 Theoretical Foundation

2.1 Fundamentals of Digital and Data Ecosystems

Jacobides et al. (2018) define a digital ecosystem as an interacting organization that is enabled by the modularity of its elements and managed without a hierarchical order. Its modular endpoints represent actors who have in common the impossibility of allocating their collective investment elsewhere (Jacobides et al. 2018). Hence, digital ecosystems are characterizable as networks of actors that are open, dynamic, and complex (Li et al. 2012; Wang 2021). The openness of the network requires a "flow of energy" between both the system and its environment and between system entities, to maintain the system state (Currie 2011). Digital ecosystems exhibit diverse temporal and spatial scales of dynamic developments, while their

complexity is subject to the number of interactions between the conjunct actors (Currie 2011; Li et al. 2012). They comprise three fundamental characteristics (Jansen et al. 2013; Oliveira et al. 2019). Their "network character" describes digital ecosystems as loosely coupled endpoints of actors. Their "platform character" implies the existence of services, tools, or technologies actors can use in the ecosystem to create value. The characteristic of "co-evolution" addresses actors mutually collaborating and connecting in the ecosystem to pool capabilities and resources for the purpose of generating innovations. The various relationships of actors to resources entail the emergence of roles comparable to functions performed by actors in the ecosystem (Hanssen and Dyba 2012; Oliveira et al. 2019). Typically, a key function emerges that is majorly responsible for ecosystem viability (Hanssen and Dyba 2012). Data ecosystems represent a subset of digital ecosystems with the purpose of sharing and jointly utilizing data (Oliveira et al. 2019). They are complex socio-technical networks that consist of, firstly, autonomous actors collaboratively utilizing data and, secondly, an environmental setting for creating, managing, and sustaining data-sharing initiatives (Oliveira et al. 2019). Well known examples are smart cities (Abu-Matar 2016), open data (Lee 2014), and scientific data communities (Lindman et al. 2015). Data ecosystems are nowadays considered an auspicious medium in our information economy to unlock the potential benefits of data across companies, industries, and entire countries (Oliveira et al. 2019). However, while data ecosystems are arguably gaining in importance, both research and practical developments of **B2C** (or C2C) data ecosystems are still in their seminal stages (Oliveira et al. 2019). This is detrimental to the information economy, given the rising importance of humans and their (personal) data for digitalization (Leidner and Tona 2021).

2.2 Data Sovereignty and Personal Data Markets

B2C data ecosystems face a multitude of legal and ethical concerns due to the systematic sharing and utilization of PD (Rantanen et al. 2019). To support legal compliance while considering ethical issues, the study uses European data law as orientation in artifact construction and the popular concept of data sovereignty as guidance in architectural design. *Sovereignty* encompasses claims to power and control that are linked to reciprocal concessions and relationships of recognition (Maritain 1950). Hummel et al. (2021) consider data sovereignty as a special form of sovereignty pertaining to empowerment of humans in terms of their data. Humans are data sovereign if they can exercise control functions over the use of their PD. This *protective claim* comprises the controllability of the entities having access to data, the determination of permitted purposes under which PD are processable, and clarity of how access and processing affect humans' data privacy and protection (Hummel et al. 2021). Among others, Hummel et al. (2021) and Lauf et al. (2022) also attribute a utilization perspective to data sovereignty. According to that *participatory claim*, self-determined PD sharing and monetization become an invariable part of data sovereignty. This inevitably requires B2C data ecosystems with data sovereignty to consider market structures, making personal data markets (PDMs) an auxiliary concept supplementing their conceptual foundation. Albeit confronted with multiple problems (Spiekermann et al. 2015), a new generation of PDMs is currently emerging in both practice (Parra-Arnau 2018) and literature (e.g., Bataineh et al. (2020), Metzger (2020)) providing a large fund of design knowledge for artifact construction. We have merged the concepts of PDMs and data ecosystems to transition the decentral data ecosystem approach from a B2B to a B2C ambit. Our artifact comprises empirically grounded design knowledge for building human-centric B2C data ecosystems with data sovereignty to share, monetize, and utilize PD.

2.3 Related Work on B2C Data Ecosystems

One of the first academic ideations of B2C data ecosystems stems from Moiso and Minerva (2012), who proposed a "user-centric" model enabling humans to control the gathering, management, use and sharing of their data. The authors framed the term "personal data ecosystem" as a data ecosystem centered around, firstly, a "Bank of Individuals' Data", secondly, providers of "personal data management services" allowing people to exploit their PD and, thirdly, the individuals themselves. However, the authors addressed a specific use case instead of designing a generally applicable model. A similar approach is represented by SOLID which currently is the best known development of B2C data ecosystems in practice. Sambra et al. (2016) describe the SOLID concept as a paradigm shift in the development of social web applications. Basically, it is a decentralized platform where humans manage their data independently of the applications that create and consume this data. As a restraint, SOLID is explicitly tailored to linked data and the semantic web, thus not generating universally valid design knowledge for B2C data ecosystems either. Rantanen and Koskinen (2020) address the question of how humans can be respected and integrated into data ecosystems. The authors state that humans should be treated as active members within those systems by giving them sufficient information and power over their data while ensuring transparency and honesty without compromising the security of both data and the ecosystem (Rantanen and Koskinen 2020). Relatedly, Koskinen et al.

(2019) proposed a governance model for people-centered data ecosystems that advocates people's rights to actively engage in the invocation and processing of PD. Both works represent very theoretical contributions that discuss ethical and societal questions rather than actually developing a solution architecture and generating design knowledge. Furthermore, related work is also embodied by projects of the Federal Republic of Germany and the EU (e.g., Kraken, dataLOFT, IDERHA) and pivotal ecosystem frameworks currently emerging in practice (e.g., Gaia-X, International Data Spaces (IDS), SOLID). However, to the best of our knowledge, developments in practice are either still in their seminal stages (i.e., SOLID) or have not yet explicitly considered the actual integration of humans with their PD (i.e., Gaia-X, IDS). Thus, we argue that, until now, the broad field of B2C data ecosystems is in a premature phase where design knowledge is scarce and both theoretical and practical progress is urgent. To that end, we define our research methodology with the purpose to design an RSA model and a set of reflective DPs for building human-centric data ecosystems with data sovereignty in B2C peripheries.

3 Research Design

We followed the design science research (DSR) paradigm, which is an accepted approach in IS research and, since our artifact is a *model* (March and Smith 1995), a suitable framework for our study (Hevner 2007; Iivari 2007; March and Storey 2008). DSR seeks to create artifacts serving organizational or human purposes by constituting an orderly structured research process for rigorously building and evaluating viable artifacts (Hevner et al. 2004; March and Smith 1995). Models, as one type of design artifact, are simplified effigies of reality accumulating specific design knowledge (March and Smith 1995). A model is an appropriate DSR artifact type for our study as we propose a representation of how human-centric B2C data ecosystems should be designed, thus exemplifying a solution statement to the prevailing problem situation described in the first two sections. Our methodological approach to DSR comprises the set of steps proposed by Peffers et al. (2007) (see Fig. 1). In line with the authors' work, those steps encompass (1) problem identification and motivation, (2) the definition of the objectives for a solution, (3) design and development, (4) demonstration, (5) evaluation, and (6) communication while following a nominal process sequence (Peffers et al. 2007). The nominal process does not determine a fixed entry point for a DSR iteration but rather allows the researchers to move forward from an arbitrary phase. The DSR process terminates once a purposeful artifact is created (Peffers et al. 2007). The final outcomes of phases 3–5 of Fig. 1 are presented in Sects. 4, 5 and 6 of this paper. For the sake of comprehensibility and due to textual limitation, we present our iteratively conducted research methods integrated into our DSR study (i.e., literature analysis, design recovery, prototyping, and expert interviews), in a compressed version. We ensure rigor and relevance of our research by strictly adhering to the DSR approach proposed by Peffers et al. (2007) for designing a novel artifact that exhibits a solution space with broad implications to the problem space prevailing in theory and practice. Following, we outline the DSR approach of Peffers et al. (2007) as contextualized in our study (see Fig. 1).

Identify Problem and Motivate: In DSR, problem identification is needed to capture the complexity, motivate research, and communicate the problem to others while enhancing the comprehensibility of the researchers' reasoning processes and their results (Peffers et al. 2007). Our motivation is to create artifacts that ease the future development of (human-centric) B2C data ecosystems and thus help to remedy the deficient situation outlined in Sect. 1. We envision an RSA (and DPs) building upon the concept of data sovereignty, thus paying tribute to societal values communicated, among others, in the European Data Strategy. However, creating these artifacts is highly complicated due to the interdisciplinarity and dynamics of B2C data ecosystems. Moreover, a lack of design knowledge about B2C data ecosystems in literature and missing instantiations in practice exacerbate our study.

Define Objectives of a Solution: The objectives of a solution are inferred from the problem definition. As far as we are concerned, there are currently no RSAs that comprise concrete design knowledge about building B2C data ecosystems. We aim to provide foundational work in this field of data ecosystem research. The solution objective is to design a technology and use case agnostic RSA that effectively solves the identified problems as well as to formulate aggregated design knowledge as DPs.

Design and Development: In this step, the RSA is developed in multiple build and evaluate cycles (Hevner et al. 2004; Peffers et al. 2007). As methods subordinated to our DSR approach, we carried out (1) a structured literature review (SLR) and (2) design recovery on analysis objects in theory and practice. In (1), we derived design implications as "quasi-requirements" from literature, whereas in (2), we retrieved design features as solutions for those implications. Our *structured literature review* consists of two separate literature analyses, the first one examining publications concerning PDMs and the second investigating literature on data ecosystems. The aim of analyzing the literature was to capture the diversity of research being conducted in the broader field of B2C data ecosystems, as delimited in Sect. 2. Examples of relevant

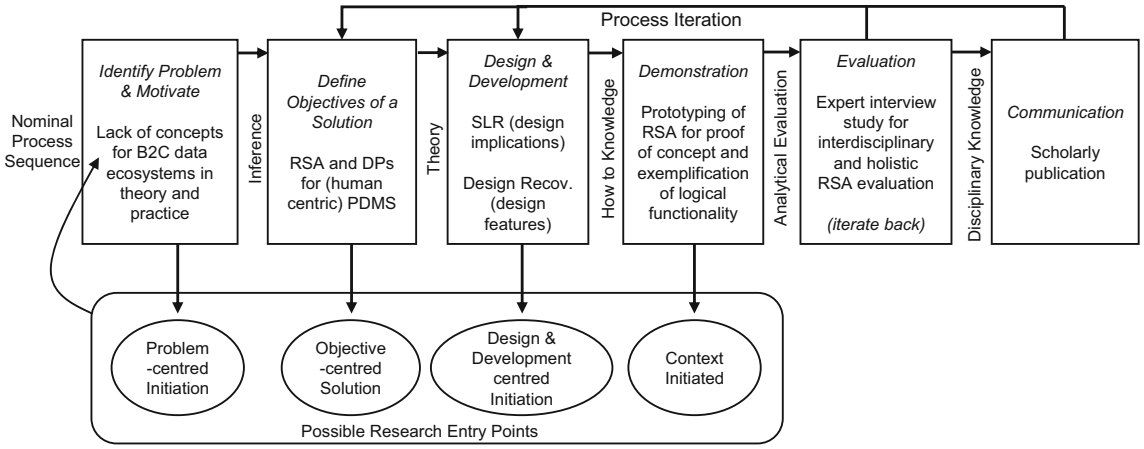

Fig. 1 DSR approach of Peffers et al. (2007) and its contextualization in the study

publications yielding design implications are Scheider et al. (2023), Kortum et al. (2022), Koskinen et al. (2019), Nagel et al. (2021), Oliveira et al. (2019), Rantanen et al. (2019), and Sambra et al. (2016). We present our SLR protocol in the first part of appendix I (available online via http://link.springer.com).

Design recovery matched our design implications from the SLR with design features extracted from analysis objects in theory (i.e., design-oriented literature) and practice (i.e., real-world solutions). We relied on Chikofsky and Cross (1990) for a methodological demarcation of our approach from other reverse engineering methods and used Biggerstaff (1989) to structure our design recovery process. While we could adopt design-oriented literature directly from the SLR (i.e., the final set of articles), a separate search for analysis objects in practice was required. We searched for the latter via the Google search engine using the incognito mode of the browser to avoid carry-over effects from previous searches and applied keywords comparable to the ones of our literature analyses. Similar to the SLR, we classified retrieved analysis objects as PDMs (e.g., Airbloc, CitizenMe, Datacoup, Datum, Kraken, Power of You, Steamr, Vetri) and data ecosystems (e.g., Gaia-X, International Data Spaces, Mobility Data Space, Resilience and Sustainability Data Space, SOLID) but also defined the category related systems (e.g., Advaneo DMP, Ocean Protocol, Own Your Data, truzzt). For details about the sources of design recovery and the classification of analysis objects from practice, we refer to the second part of online appendix I. From our retrieved analysis objects (i.e., design-oriented literature and real-world solutions), we extracted design features for artifact development. We constructed or synthesized implementation-agnostic abstractions in a way that exhibited an inherent reference character (Chikofsky and Cross 1990). In our study, we define design features as specific practices in which certain functionalities are integrated into analysis objects. While our finally developed DPs abstract from technical or conceptual specifics, design features close this last step of conceptualization (Meth et al. 2015). By means of design recovery, we developed our RSA iteratively by logically and functionally merging extracted design features on higher levels of abstraction (Asif 2003). This process was guided by prototyping (demonstration), expert interviews (evaluation), and discussions in the research team, which were all particularly useful to choose best practices if ambiguity occurred among extracted design features. The ultimate results of the design and development phase are presented in Sect. 4.

Demonstration: The efficacy of the artifact to solve the problem must be demonstrated through the artifact's context initiation (Peffers et al. 2007). Since data ecosystems cannot be entirely implemented within an ordinary research project due to their decentral nature, we mainly used prototyping to provide a proof of concept and to evaluate design elements developed in the previous phase. Even though technical feasibility is arguably inherent in large parts of our conceptual artifact due to rigor reliance on the described construction methods (i.e., design recovery), our prototypical instantiation is an initial proof of concept to be extended in future work. For *Frontend* prototyping, we ideated platform-agnostic human user interfaces based on constantly refined versions of our RSA evolving in ongoing DSR iterations. This served to identify central interactions between the artifact's partial models, resulting in their continuous enhancement. We identified supplementary design implications and meaningful ways for their integration into the artifact, improving both its comprehensibility and logical structure. In the *Backend*, we limited our proof of concept to the design of information exchanges and storage between artificial subsystems (see Fig. 2). We used a virtual machine as a testbed environment and Docker containers to simulate decentral endpoints of the conceptualized ecosystem. The connections between

containers are implemented via Vagrant to emulate their interactions in the RSA (see Sect. 5).

Evaluation: We applied expert interviews to measure how well the artifact supports a solution to the problem. This activity involves comparing the objectives of a solution to actually observed results from artifact demonstration (March and Smith 1995). However, in the interviews, we evaluated our latest results from both artifact demonstration and the design and development phase, i.e., the RSA. This is reasonable as the RSA was designed to enable an easy elaboration on the functionality underlying the artifact (see Sect. 6 and online appendix II). Thus, our interviewed experts were facilitated in the artifact evaluation with respect to the solution objectives and its overall comprehensibility (Sonnenberg and Brocke 2012). At the end of an evaluation activity, we decided whether to iterate back to the design and development phase trying to improve the artifact's effectiveness or to terminate the DSR process, leaving further improvement to subsequent research. The evaluation phase is summarized in Sect. 6.

Communication: The problem and its importance are communicated to relevant audiences along with the artifact, its utility and novelty, the rigor of its design, and its effectiveness (Peffers et al. 2007).

4 Reference System Architecture for Human-Centric B2C Data Ecosystems

In the following, we describe the RSA at a level of abstraction that facilitates the readers' understanding while also linking to the reflective DPs inferred from the architecture. In the online appendix II, we give full particulars of our modified 3 + 1 View Model derived from Kruchten (1995) and Reidt (2018), serving as the justificatory knowledge which both informs our design (Gregor and Jones 2007) and structures the RSA. Cumulatively, these "views" on the RSA aggregate the entirety of architectural design knowledge gained about human-centric B2C data ecosystems in the course of our study. They comprise a functional view, a role model, a distribution view, and a process view. While online appendix II still provides a compressed version of the RSA, its detailed architectural design knowledge can be provided on request (i.e., UML sequence diagrams with comprehensive explanations). The same holds for our prototype built upon this architectural design knowledge (see Sect. 5).

The RSA describes a human-centric B2C data ecosystem consisting of *data suppliers* (i.e., Data Owners, Data Providers) who share data at large-scale with *data demanders* (i.e., Data Consumers, Workbench Providers). The latter processes and utilizes that data within the ecosystem. In between, *intermediary actors* (i.e., Broker, Data Quality Curator, Registrar, Fiduciary, Vocabulary Curator) provide services to enable data sharing, monetization, and utilization based on European data law, while ensuring essential model properties such as transparency, fairness, and trust among actors. The composition of actors implies a decentralized infrastructure as the technical foundation of our RSA. Within this decentralized infrastructure (see Fig. 2; [a]–[x]), Data Owners (human data subjects) and Data Consumers exchange private datasets by means of their interface-providing actors [a], i.e., Data Providers and Workbench Providers. Data are imported [b] and stored [c] by Data Providers who offer *Personal Data Storages* (PDS) as decentral storage locations to Data Owners. Following the idea of SOLID, multiple PDS are connected to and maintained by a Data Provider's *Data Resource Port*. Data Owners have permanent access to data via their PDS accounts, where they can manage stored data sovereignly, i.e., asserting data subject rights [d] (Art. 12 GDPR et seqq.; see Fig. 6). However, due to this strong orientation of our RSA toward data sovereignty and human-centricity, the model requires Data Owners to engage with the system to an extend probably overcharging their processing capacities and convenience (Bester et al. 2016; Scheider et al. 2023). Thus, we integrated automation mechanisms for the many obligations and tasks of the Data Owner. Specifically, the RSA conceptualizes a hybrid consent model allowing the Data Owner to shift entire "activity types" (e.g., importing, storing, and offering data), for particular data classes, by means of broad consent [e], to the Data Provider (see Fig. 4). Under the aegis of legal experts (see Sect. 6), the RSA has been designed to allow the application of broad consents in all but one of our defined activity types (i.e., responding to data orders [f]).

In the RSA, data are imported [b] and stored [c] either on individual user requests [d] or automated by the Data Provider [e]. During import, data are standardized, and metadata are extracted in the Data Resource Port that leverages data class specific data models. Important vocabularies (i.e., data models [g], usage policy ontologies [h]) are maintained in the *Vocabulary Catalogue* and provided on request, supporting semantic consistency. Each dataset receives a unique metadata ID which is a *sine qua non* for all communication between actors in the system. To store imported data in a PDS connected to the port, a usage policy (see Fig. 7) must be specified by either the Data Owner [d] or Provider [e]. Usage policies serve as input variables in the *License Repository* [i] that returns a set of standardized licenses [j] exhibiting the best fit to the input policy by approximation. Data licenses are well-defined data contract terms comprising both a machine and human readable form [e.g., Governatori et al. (2013)]. They represent meaningful sets of usage policies and are maintained in the License Repository by the Registrar. In the

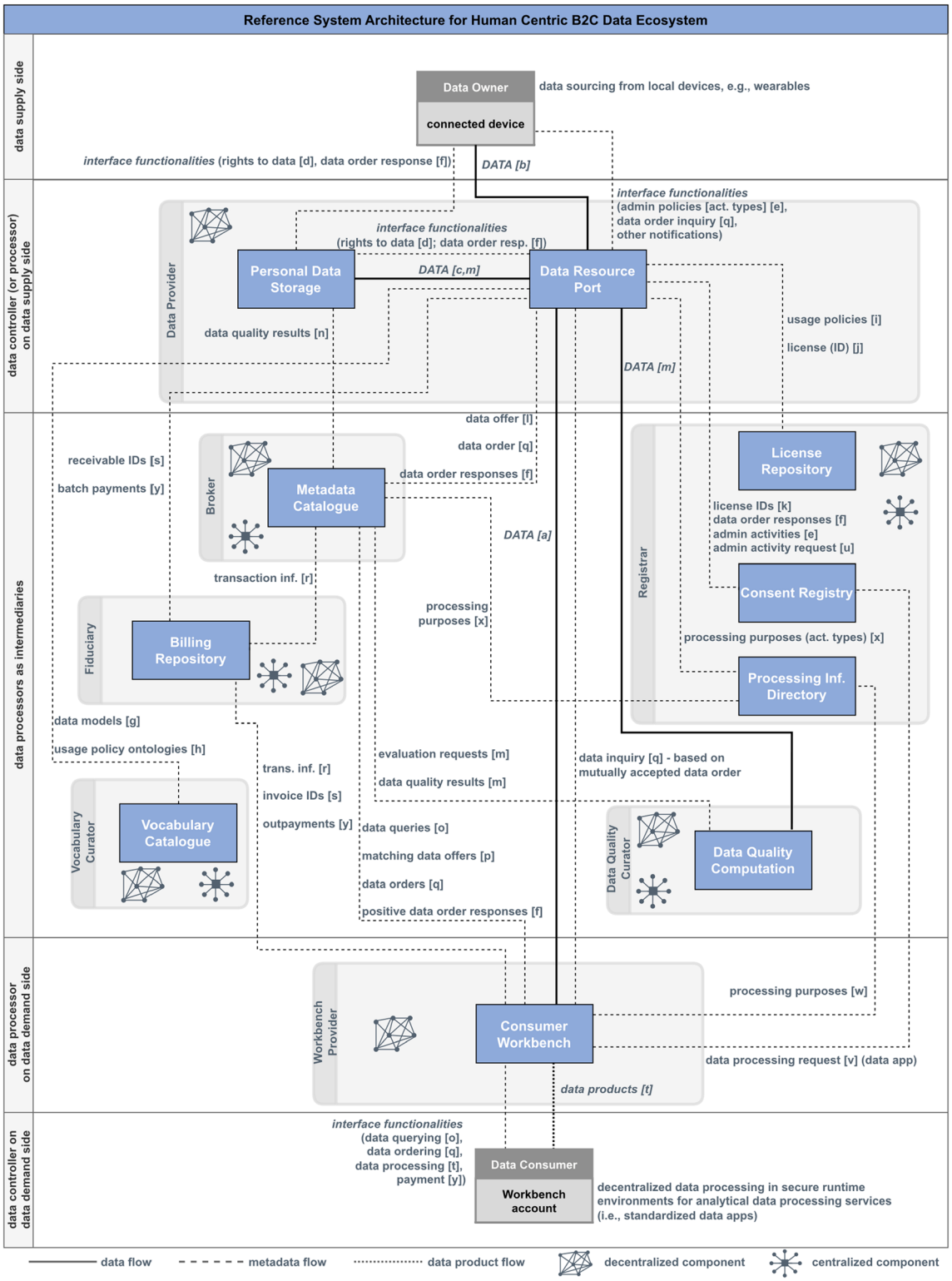

Fig. 2 Reference system architecture

RSA, the standardization of data usage policies is crucial to find an equilibrium between the necessities of data governance (i.e., technically enforceable, human-centric usage control) and data economics (i.e., effective, efficient PD processing), which is a binding condition for viable B2C data ecosystems (Scheider et al. 2023). The enablement of

Data Owners to formulate a usage policy upfront, instead of selecting licenses directly, serves to minimize the risk of nudging effects potentially curtailing data sovereignty (Lauf et al. 2022). Upon the Data Owner [d] or Provider [e] selects a license, the metadata ID of the dataset is linked to the respective license ID. Such ID pairs are stored by the Registrar in the *Consent Registry* [k], where all kinds of consents arising in the system (i.e., data licenses [k], data order responses [f], admin policies [e]) are recorded consistently and transparently. These consent records are crucial to evaluate data processing requests and, in retrospect, proof the legitimacy of processing activities.

To both enable and incentivize data sharing based on data licenses, metadata of stored datasets can be published as data offers via the Broker in the *Metadata Catalogue* [l], integrating a market mechanism. To this end, datasets receive a quality evaluation conducted by Data Quality Curators in the *Data Quality Wiki* [m]. Similar to license IDs, computed quality scores are attached as an extension of a dataset's metadata in the PDS [n]. Given an existing data quality score and license ID, a data offer corresponding to a dataset's metadata is published by the Broker in the Metadata Catalogue. Data offers can be queried by Data Consumers [o]. These actors generate data orders [q] from (a subset of) the queried matching set [p]. To this end, they specify the processing purpose of the data order and a price for each dataset inquired. For Data Consumers, a naïve approach to price determination is to distribute the pecuniary value expected from processing inquired PD (i.e., the data product) to the total number of the datasets contained in the data order. In the RSA, data orders are distributed by the Metadata Catalogue to the Data Resource Ports to which the metadata IDs of datasets, inquired by the order, are linking [q]. If the Data Owner consents to a data order [f] received via its PDS account [q] (see Fig. 5), data are shared [a], and the data order's transaction meta-information are transmitted by both the Broker and the data receiving actor to the *Billing Repository* [r]. The component is operated by the Fiduciary, who executes and documents payments based on the transaction metadata received. This particularly encompasses generating, recording, and distributing receivables and invoices for shared data to Data Providers and Consumers [s] based on the prices indicated in the data order, thus enabling payments [y]. Exchanged data are stored during their subscription time (or until revocation by the Data Owner) in decentralized *Workbenches*, maintained by Workbench Providers. They allow Data Consumers to execute data apps, generating data products [t]. While the particular processing activities performed by data apps do not require any specification in the RSA, each data app must be directly linked to a processing purpose (see Fig. 7 or online appendix III for very abstract examples). Importantly, the RSA prescribes to well-define all data processing purposes possibly occurring in the system. Purposes embody an integral part of data licenses (data supply side processing authorizations) as well as data apps and data orders (data demand side processing requests). Their well-definition and limitation enable human-centric data governance (see online appendix III and Sects. 5, 7), which implies impermeable system boundaries prohibiting the export of content data on the system's data demand side. Thus, Data Consumers can only access the results of processing activities (i.e., data products) that are generated on their behalf by Workbench Providers (i.e., data apps).

The Registrar maintains the Consent Registry ex ante evaluating data processing requests of the Data Providers [u] and Consumers [v] by leveraging stored consent records (i.e., data licenses and admin policies; see above). Depending in the inquiring actor, the Registrar evaluates processing requests for a dataset based on either: (1) data licenses authorizing, among others, well-defined processing purposes and (groups of) Data Consumers [k] (see Fig. 7); or (2) admin policies permitting the conduction of activity types for a particular Data Provider [e] (see Fig. 4). Essentially, for all data processing activities that were both authorized and actually performed, the responsible actor records, in the *Processing Information Directory*, its own identity (i.e., Data Provider/Consumer; Art. 4 Nr. 7 GDPR) and the processing purpose [w], respectively activity type [x], given the exact same entry does not yet exist. This provides transparency of data processing. Together with the Data Quality Curator [m], the role of the Registrar is crucial in the RSA to integrate trust mechanisms which, in our investigations, appeared as a fundamental design element for data ecosystems generally, e.g., Nagel et al. (2021).

Figure 2 depicts the RSA on the highest level of abstraction aggregating the entire 3 + 1 View Model. The shortcuts [a]–[x] relate to the interactions shown in Fig. 2. The corresponding processes can be provided with detailed descriptions on request. Moreover, Fig. 2 indicates which system components may be distributed, centralized, or both (see icons). The level of decentralization depends mainly on the (de-) centralization of intermediaries. Decentralized intermediaries increase system complexity since an orchestrating meta-level is required, e.g., a catalog of Metadata Catalogues, a repository of Consent Repositories, et cetera. For the sake of simplicity, we assume centralized intermediary system components/roles in our RSA.

5 Artifact Demonstration

March and Storey (2008) and Sonnenberg and Brocke (2012) recognize prototyping as an important method to

evaluate DSR artifacts. Thus, in our design process aligned to Peffers et al. (2007), artifact demonstration represents, at the same time, the non-theoretical supplement of our artifact's evaluation (see Fig. 1). In the following, we present an excerpt from our prototype to amplify the logical functionality of the RSA, facilitating an understanding of its complex content presented in Sect. 4. We emphasize the Data Owner perspective to accentuate human-centricity and data sovereignty in artifact design. By means of prototyping, we extended our knowledge about the artifact in general. In particular, we evaluated our RSA in terms of additional design features (e.g., the design of the Data Owner interface) and effective possibilities for instantiation. Examples are that (1) *Java* is recommendable to set up our decentralized system consisting of different servers that are operated via REST interfaces. (2) *Openfeign* is an eligible tool for simplifying information exchanges between the conceptualized system components. (3) In implementation, these *system components* should be integrable and interchangeable as to scalability and flexibility. (4) *Eureka* is suited to provide an overview of the prototypical instantiation whose characteristics imply high complexity (see Sect. 4). In the following artifact demonstration, the shortcuts *V1–V28* refer to the respective views of the Data Owner frontend *click-dummy* provided in online appendix IV (i.e., the slides of the executable PDF file). They provide a visualization of the presented content beyond Figs. 3, 4, 5, 6 and 7 to further support the understanding of readers.

From the Data Owner perspective, our prototype contains the here addressed functions which were derived from the RSA described in Sect. 4. After login to the system (see *V1*), the Data Owners access their Digital Me enabling them to enter the Data Resource Ports (DRP) of all Data Providers at whom these users are registered and have stored any personal data (see *V3*). Following the SOLID concept, one can assume, that in our modelled system, Data Owners are likely to store data at multiple Data Providers who offer PDS with diverse characteristics (see *V25*), thus being adequate for different data classes (Sambra et al. 2016). For example, a PDS emphasizing security rather than usability may be suitable for health data. The Data Owner selects a Data Provider entering *MY Data Resource Port* to access stored data. Recalling Sect. 4, a provider's DRP connects to its Personal Data Storages, at which Data Owners maintain accounts and manage their stored data, either manually or automatized via admin policies. Figure 3 shows the navigation paths branching off from the DRP's Data Owner frontend.

Firstly, the Data Owner can navigate to *MY Billing Repository* (see *V5*) via the homonymous button to access the transaction meta-information (i.e., receivables) of executed data orders (e.g., payment status, price, and the metadata IDs of datasets). Secondly, *MY Processing Information Directory* (PID) lists processing logs of all data stored at the selected Data Provider that were processed in the system based on well-defined purposes (activity types or Data Consumer purposes; see *V20*). Following our legal experts' interpretations of Art. 30 GDPR, recording the purpose of data processing is sufficient. The PID stores no duplicates. Thirdly, the Data Owner can access *MY Consent Registry* (see *V6*) containing three sub-registers that record specific kinds of consents associated with the selected Data Provider: (1) the "admin policies" for activity types the Data Owner granted to this provider (see *V7*); (2) the "license IDs" attached to

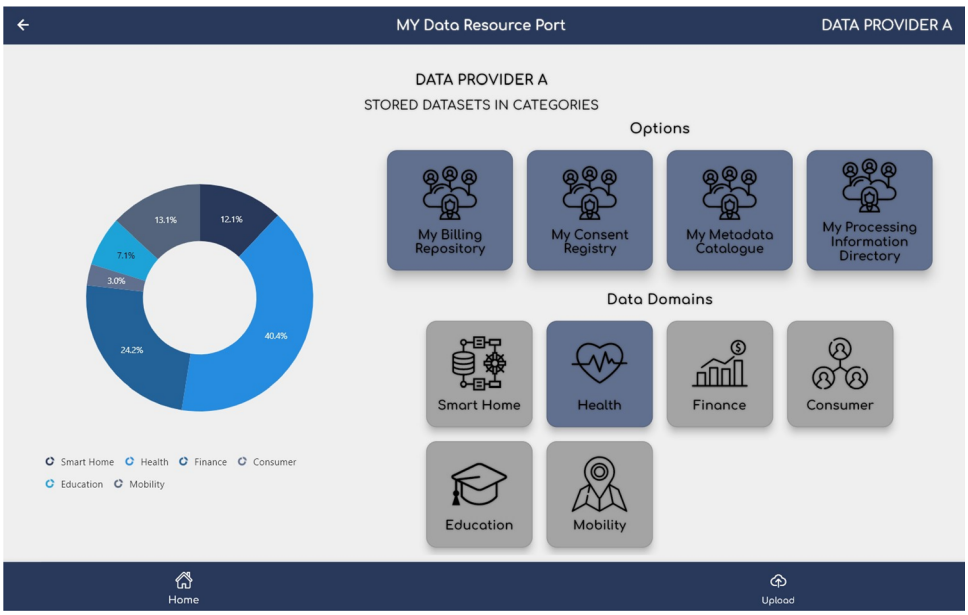

Fig. 3 MY data resource port

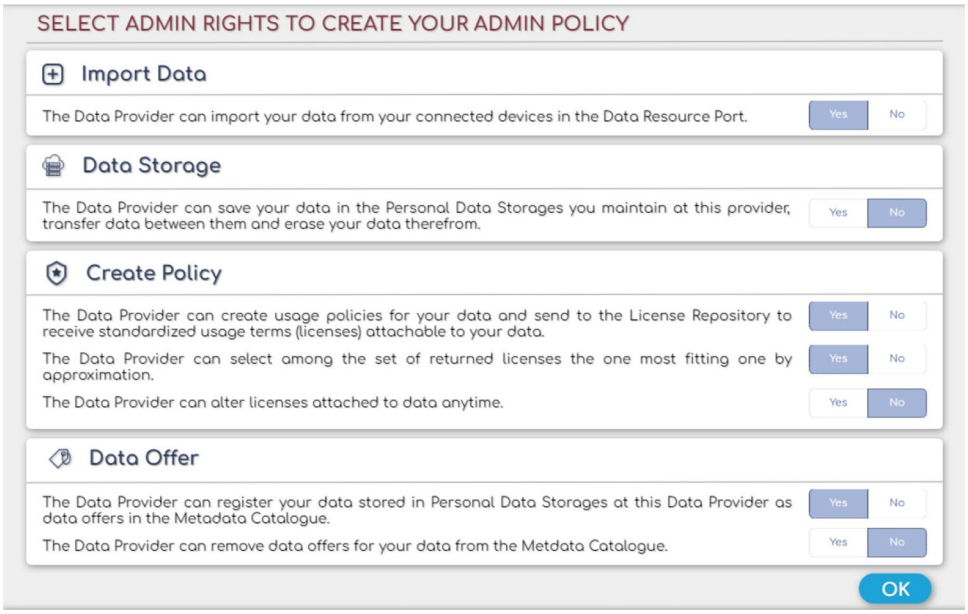

Fig. 4 Admin rights inferred from activity types for admin policy creation

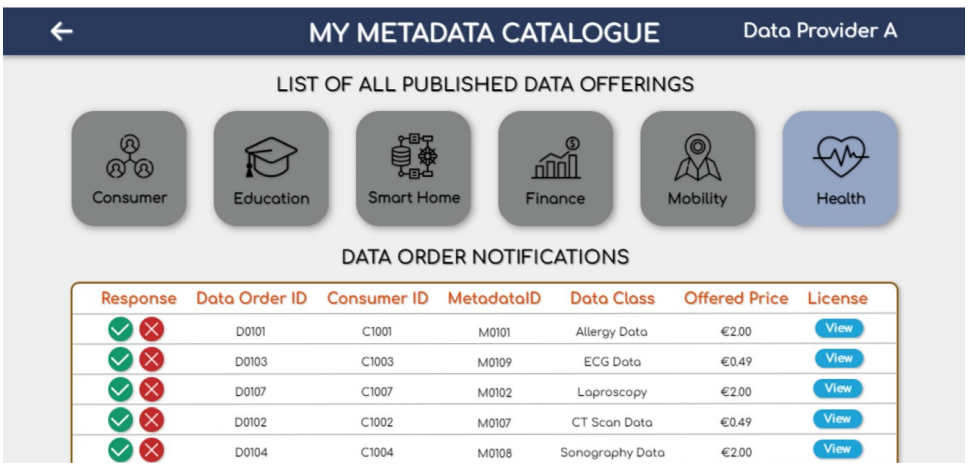

Fig. 5 MY metadata catalogue

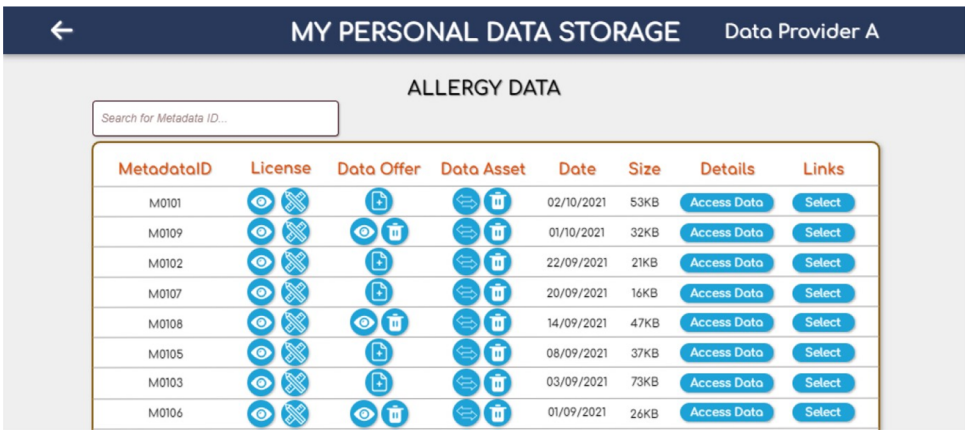

Fig. 6 MY personal data storage

datasets stored at this provider (see *V12*), and the "data order responses" given by the Data Owner for datasets stored at this provider which were inquired by Data Consumers (see *V14*).

To create admin policies, the Data Owner navigates via *MY Consent Registry* to *MY Admin Policies*, as addressed by (1). Since admin policies are defined directly for the Data Provider who shall receive the admin rights (i.e., set

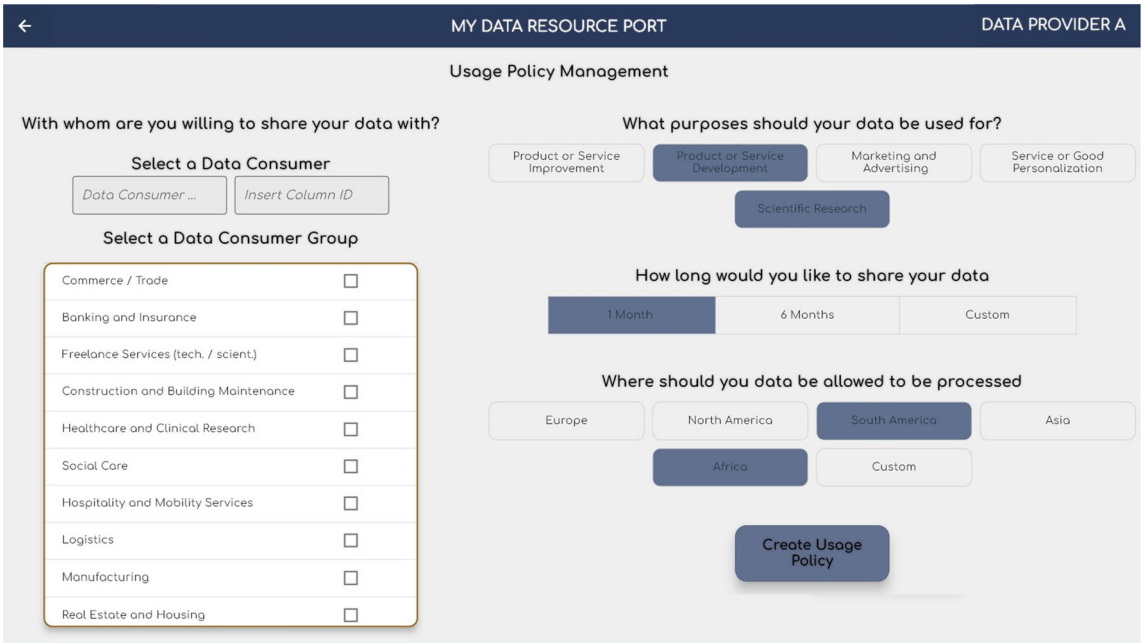

Fig. 7 Usage policy management (abstract example)

of authorized activity types), the Data Owner can start by determining the data classes for which the admin policy applies. Next, the Data Owner specifies the particular admin rights constituting the admin policy (see *V7–V11*). Figure 4 visualizes our admin rights (i.e., Data Provider processing purposes) that are directly inferable from the activity types conceptualized in the RSA. In particular, we envision admin rights grantable by means of broad consent for activity types associated with importing and storing data, specifying usage policies, determining licenses as well as managing data offers, excluding data order responses.

Recalling Fig. 3, Data Owners can also navigate from *MY Data Resource Port* to the data offers they maintain via the selected Data Provider at the Metadata Catalogue (see Fig. 5). Among others, *MY Metadata Catalogue* allows Data Owners to give specific consent or refusal to data orders. In the RSA, this is the only activity that cannot be automatized by admin policies, as sufficiently tight appropriation is considered impossible (see Sect. 6). The Data Owner can access all data offers maintained via the selected Data Provider in diverse data domains. For example, *MY Health Data Offerings* (see *V18*) allows observing those instances' metadata, e.g., the number of data orders having requested a particular offer, the size of the underlying dataset, its attached license ID and identifier (metadata ID), as well as the results of quality evaluation. Data offers can be generated and deleted either automatized by admin policies or manually via *MY Personal Data Storage* (see Fig. 6). *MY Personal Data Storage* is reachable by following one of the paths branching off from the DRP interface that are classified as "data domain" (see Fig. 3). These paths are crucial as they allow the Data Owner to navigate to the actual PDS endpoints. To this end, the user selects a data domain and, as a consecutive step, a data class (see *V21*) leading, by default, to the corresponding *MY Personal Data Storage* location (see Fig. 6). In the RSA, such endpoints exist for all data classes a Data Owner stores datasets in a PDS of a Data Provider.

By means of the functions provided in a PDS (or other components), Data Owners can always manually assert their rights to data (Art. 15 et seq. GDPR). These rights are derived from the GDPR, adopted as dedicated functions in the RSA, and visualized in the prototypical Data Owner frontend, e.g., the buttons of Fig. 6 (i.e., erasing, moving, offering, and accessing data as well as determining a data license).

To import a dataset and for storing it in a PDS, either the Data Owner or the Data Provider uploads data from a user device connected to the DRP (see *V24*), chooses a PDS to which the data should be transferred (see *V25*), specifies an initial data usage policy (see Fig. 7), and selects one of the licenses proposed by the system based on the usage policy defined (see *V27*). Since our RSA is a conceptual model with empirical grounding, technical implementations beyond the prototype are subject to future research. Due to the intricate deployment of the prototype (e.g., several server instances, required Java runtime environment), we only present our fast intelligible Data Owner frontend click-dummy in the online appendix IV.

6 Artifact Evaluation

In DSR, evaluation represents a central methodological component and significantly influences the final artifact (Hevner et al. 2004). March and Smith (1995) refer to evaluation in design science as a second element within the DSR process. Our evaluation's objective was to prove that the RSA addresses our RQs, complements existing domain knowledge, and represents a suitable artifact for facilitating the construction of (human-centric) B2C data ecosystems. Since prototyping already served as a more "technical" evaluation (March and Storey 2008), we used extensive expert interviews as an additional qualitative evaluation method carried out within design iterations. Thus, similar to prototyping, received feedback could directly be considered in subsequent construction cycles (see Fig. 1).

We conducted semi-structured interviews to validate our artifact's feasibility and understandability, as those are typical evaluation criteria for DSR artifacts (Sonnenberg and Brocke 2012). Given the interdisciplinarity of the field, we chose our interviewees by ensuring expertise in the four dimensions postulated by Meister and Otto (2019). Consequently, the selected experts stem from diverse areas (e.g., researchers, lawyers, and data scientists) but are all skilled with regard to data ecosystems. Hence, they have both thorough knowledge and sufficiently differentiated points of view on the topic to give extensive feedback. We followed an expert sampling approach, inviting industrial and scientific partners from research projects and our personal networks (Bhattacherjee 2012). In the run-up to the interview, experts received a detailed but compact summary of the artifact. To ensure that this summary adequately represents the RSA, we made use of aggregated contents of its partial models (see online appendix II), which were presented to experts depending on their profession and technical experience. These aggregated contents particularly encompassed the UML sequence diagrams of the Process View (to be provided on request), the module table of the Functional View (see Table 1 in online appendix II), the synopsis of the actors' roles and obligations defined in the Role Model (see Tables 2–4 in online appendix II), and the holistic RSA overview associated with the Distribution View (see Fig. 2). By relying on this consolidated information in our artifact presentation, we ensured that experts understood the purpose and the design elements of the RSA and thus could contribute a meaningful evaluation. For technically-versed experts (#I-III, VII), we mainly mediated content by means of our UML sequence diagrams. Concerning less technically experienced experts (#IV-VI), we emphasized the other kinds of consolidated information to provide an intelligible summary of the RSA. Interviews were conducted by one researcher, whereas usually two researchers were responsible for coding and deriving implications for the subsequent design iteration. Since concrete design elements were evaluated by experts, specific design issues could be addressed resulting in the confirmation of existing or the suggestion of new design features. Specifically, a *design issue* is a problem in association with a design element of the presented RSA for which the research team either evaluates an existing solution (i.e., experts confirm a design feature) or asks for a problem solving approach (i.e., experts propose new design features). Naturally, during the presentation of the RSA, experts also encountered yet unknown design issues and suggested helpful design features. However, regardless of their origin, design issues could be addressed by experts directly: "How can the issue be solved in the corresponding (or another) design element?". Table 1 exemplifies some design issues, elements, and features on a high level of abstraction to foster the understanding of our artifact evaluation method by means of expert interviews.

As a result of this evaluation approach, we received substantial feedback leaving hardly any space for false interpretation, which allowed us to dispense with detailed coding. Using Strauss and Corbin (1990) as a very rough orientation, we extracted quotes (Pratt 2008) directly addressing one or more design implications for the artifact (open coding), generated a code as a further aggregation of a set of quotes (axial coding), and classified codes as RSA specific descriptions of design features (selective coding) to be integrated in the next design iteration. On the level of codes, we permanently checked for conflicts. Due to the concreteness of addressed design elements, and the unambiguity of received feedback, we argue for the reliability of our coding and, thus, for a minor relevance of subjectivity issues. If generated codes showed ambiguity, they were discussed with the research team until a consensus was reached. In the online appendix V, we give examples of quotes and coding. In the interviews, experts focused on their professional focal points stated in the third column of Table 2. They positively evaluated the artifact's technical feasibility (#I–III, VII), comprehensibility (#I–VII), legal compliance (#IV-VI), and effectiveness in achieving RQ1 (#I–VII). This particularly holds for our hybrid consent model based on admin rights (#IV-VII). Following, we present some excerpts of the qualitative expert interview results.

The experts assured that our RSA considers *data subject rights* according to Art. 15 GDPR et seq. (#VI), whereby they in particular focused on the following ones. The rights of access and deletion through revocation are integrated into Personal Data Storages (see online appendix II and Fig. 6). Data portability, required by Art. 20 GDPR, is supported by the Vocabulary Catalogue, that provides data models for standardization (#V). The Consent Registry and

Table 1 Examples of design issues, elements, and features (confirmed [c] or proposed [p])

Design issue of	Design element solved by	[c]/[p] Design feature
"Can the market mechanism determine meaningful prices for data?"	Market mechanism	[c] attributing pricing power to data consumers
"How can we leverage broad consents in our consent model most effectively?"	Hybrid consent model	[p] designing activity types to maximize broad consents
"How can PD be processed effectively given our human-centric usage control?"	Data governance structure	[p] standarding data usage policies through licenses
"How can we minimize the documentation effort related to PD processing?"	Trust mechanisms	[p] recording consent proofs and processing purposes
"Does the RSA provide sufficient technical and organizational safeguards for PD processing?"	Decentralized infrastructure and trust mechanisms	[c] executing data apps in secure runtime environment

Table 2 Table of experts listed according to the domain, role, field, and duration of the interview

Experts (#I-VI)	Role	Field	[min]
#I: Science	Scientist	Data ecosystems, IS engineering	97:45
#II: Industry	Data Scientist	Data analytics, Quantitative modelling	86:05
#III: Science	Professor, Scientist	Digital health, Data ecosystems	70:52
#IV: Science	Lawyer	Data law, Data trusts, PIMS	50:16
#V: Industry	Lawyer	Data law, Data privacy, Data trusts	67:52
#VI: Industry	Ethicist, Lawyer	Data law, Ethics	39:36
#VII: Science	Scientist	Consent modelling, IS engineering	79:56

the Processing Information Directory provide evidence in terms of processing purposes of different actors and their legitimacy, satisfying Art. 6 and 30 GDPR. This is particularly important for the Data Provider and Data Consumer, as controllers of PD, who are obligated to prove that they have adhered to declared consent and appropriation in data processing (#IV-VI). Since Art. 30 GDPR does not precisely state the required characteristics of documentation, one must determine the minimum requirements arising in the specific system contexts to satisfy obligations ("risk-based approach"). Art. 30 GDPR implies to record the consent and purpose of data processing but not the applied means (i.e., processing activities). Thus, indicating what data were processed by actors and their purposes is sufficient (#V). Furthermore, PD cannot be *licensed* in contract law, as the data subject must always be able to revoke consent (#IV–VI). Hence, in the RSA, the Data Owner can withdraw consent to a data order during its underlying license's lifetime at any time, instructing the Workbench Provider to delete the data immediately. However, such consent revocation leads to the Data Owner losing the claim to the mutually agreed price. This uncertainty for Data Consumers cannot be eliminated (#V). However, the RSA provides an incentive as Data Owners are only paid if consent has not been withdrawn. Moreover, *admin policies* for Data Providers are fundamentally problematic since such policies are inevitably formulated in broad terms, thus making a strict appropriation of data processing difficult or even unlikely (#IV–VI). As a result, there is a severe risk that consent requirements (Art. 7 GDPR) are circumvented (un-) consciously. In principle, those necessities must be imposed on admin policies in the sense of an advance displacement (#V). The Data Provider must inform the Data Owner about the consequences of an admin policy, which must only be grantable for the most specific purposes possible (i.e., activity types) and revocable anytime. Design implications of such admin policies, especially the required scope of their appropriation, are not legally regulated and represent a grey zone in European data law (#IV, V). Admin policies might be implementable through meta consent models (Ploug and Holm 2016), which are applicable if data processing comprises a low risk of abuse (#V), e.g., recital 33 GDPR for research. This reasoning can be transferred to our model if the Data Provider has a justifiable interest in the welfare of Data Owners (#V). Given the unclear legal situation in Europe, we envisioned a hybrid approach in the RSA and ideated, under the aegis of legal experts (#IV–VI), admin policies for all activity types for which "sufficiently tight" appropriation is assumed possible (see Fig. 4). As described in Sect. 4, admin policies recorded in the Consent Registry are checked by the Registrar against processing requests of the Data Provider before their execution. An intuitive

graphical Data Owner interface to manage admin policies is of utmost importance (#IV; see Sect. 5).

Besides legal topics, experts mentioned that *data apps* might be problematic as Data Consumers' access to the actual data is forbidden. Data scientists frequently require such access for model validation, own quality checks, and individualized analyses (#II, III). However, for our RSA, they recognized data apps as currently the best feasible approach. Furthermore, the *process of data ordering* was addressed, specifically, the problem of skewness in generated matching sets. The selection of attributes in data search queries can lead to distribution biases in matching sets due to a subconscious and unintended focus on certain segments (#II). An expert argued that a filter for high data quality, provided by a query functionality in the Metadata Catalogue, is likely to correlate with a higher economic status of Data Owners. As a proposed solution, the provision of information about the distribution of metadata in matching sets was suggested (i.e., price and quality distribution histograms), as well as a function to randomly select a subset of the entire matching set for the inclusion in the data order to enhance representativeness (#II). In terms of our approach to *data pricing by the consumer*, experts stated that data sovereignty would be strengthened by empowering Data Owners to actively determine prices themselves rather than assuming a passive role (#I–III). However, they acknowledged that for such purposes, Data Owners would need much more supportive information in a format not exceeding their processing capacities (#II, III). Examples were to display, for comparable datasets, current market prices traded, contemplating the idea of a "data stock exchange" or indicating (averages of) prices determined by other Data Owners recently in the system (#II). The experts also recognized that an alternative *pricing by the owner* mechanism might be, if viable, highly cumbersome and user-unfriendly (#I–III). Lastly, several experts mentioned the idea of *returning data products* generated by means of data processing not only to Data Consumers but also, at least in certain parts, to Data Owners (#I, III, VI). This two-sided business model provides space for future work far beyond the scope of this DSR project. Overall, given its complex system context, the experts ascribed the RSA an adequate level of comprehensibility, adaptability, and practicability (#I–VII). They also accentuated the suitability of our modified 3 + 1 View Model (i.e., the consolidated information) to facilitate complex content (see online appendix II; #II, VII).

Hennink et al. (2017) propose parameters for assessing saturation in qualitative interviews and estimating appropriate sample sizes. These parameters encompass (1) the *study purpose* (i.e., broad vs. narrow issues); (2) the *study population* (i.e., the number and heterogeneity of relevant attributes required); (3) the *sampling strategy* (i.e., iterative vs. fixed sampling); (4) the *data quality* (i.e., "thick" vs. "thin" data in terms of the deepness and richness of inferable insights), (5) the *type of codes* (i.e., explicit and concrete vs. conceptual); (6) the complexity and stability of the *codebook*; as well as (7) the *saturation* of the *objective* (i.e., core codes or all data) and the *focus* (i.e., "code" or "meaning" saturation). Both, an adequate sample size and the saturation of results are determined by the combined influence of all parameters rather than a single parameter alone (Hennink et al. 2017). Even though the RSA is complex in itself, the experts only addressed specific design elements based on their professions (see above; Table 1). Thus, they evaluated certain thematic issues in particular instead of explaining a complex phenomenon generally (1). Furthermore, our interviewees exhibit heterogeneity in their concrete disciplines (i.e., legal, economics, technology), but their population is homogenous in terms of expertise about data ecosystems as the main matter of subject (2). For recruitment, we applied fixed sampling as the experts were chosen to ensure an adequate distribution with regard to the aforementioned disciplines (3). Due to the exceptional expertise of all interviewees in their disciplines, we collected high quality data providing valuable insights (i.e., specific design features for design issues; (4). Likewise, we generated explicit and concrete types of codes (5). The complexity of the codebook was low since particular design elements were assessed by experts with respective knowledge (6). Finally, the goal of the interviews was to achieve code saturation (7), which we define as the justified claim that all potential design issues pertinent to our RSA have been solved and no further issues can be encountered (at the given level of model abstraction). Since (1)–(7) all suggest a rather low sample size as sufficient and each design issue was solved after interviewing six experts, we argue based on Hennink et al. (2017) that a sufficient saturation in results was achieved after seven extensive expert interviews (see Table 2).

7 Formalizing Design Knowledge for Human-Centric B2C Data Ecosystems

Summarizing what we have learned, we formalize the design knowledge prevailingly as design principles (DP). We pursue "reflecting upon what has been done" (Gregor 2009) and codify relevant design knowledge as DPs following the template of Chandra et al. (2015). This reflective process of codification entails the significant advantage of making what we have learned available to others at a different point in time in easily actionable prescriptions for action (Cohendet and Meyer-Krahmer 2001). We used the method of Möller et al. (2020) to elicit reflective DPs based

on design knowledge gained in artifact construction. Below, we present our reflective DPs for human-centric B2C data ecosystems that represent a model-specific further development of the supportive DPs proposed by Scheider et al. (2023). Consequently, we provide an empirically justified answer to RQ2.

DP1: *Provide a market mechanism in B2C data ecosystems to support their economic viability and to assure the alienability and the excludability of the data traded therein, while incentivizing systematic data sharing, data monetization, and data utilization processes.* **Rationale**: B2C data ecosystems must transform PD from a common pool resource into a private good permitting the integration of a market mechanism. Such mechanisms can be instantiated by (meta-) data catalogues enabling processes to share and monetize data within the B2C data ecosystem. For data consumers, the market mechanism facilitates to buy or, more realistically, subscribe to data. Therefore, it must contain adequate query and filter functionalities as well as enable efficient matchmaking between the data supply and demand side. Importantly, the market mechanism must circumnavigate the need for data consumers to access the content data. To this end, it may disclose appropriate metadata about datasets offered. Moreover, the market mechanism must support fast data access. A prevailing problem in that regard is to determine an objectively fair price of PD. A naive solution is data consumers leveraging the metadata about datasets disclosed by the market mechanism to estimate a pecuniary value from data processing. As a consequence, they infer a willingness to buy that allows to formulate prices for individual datasets. For data subjects, the market mechanism provides an incentivization to engage in data sharing within the system because they receive a share of the economic exploitation potential yielded from processing their data. Importantly, data subjects must always be in charge of all data sharing activities and able to revoke corresponding consents immediately. In principle, market mechanisms integrated in (human-centric) B2C data ecosystems must always undergo an examination with regard to whether humans' data sovereignty might be curtailed. An arising problem is individuals' lacking digital literacy, as they are hardly aware of the economic value linked to their data (Spiekermann et al. 2015).

DP2: *Provide a hybrid consent mechanism in B2C data ecosystems as the enabler of usability and automation, allowing data subjects to shift required manual interactions with the system to a deputy actor, while the scope of the attributable interactions depends on the applicable jurisdiction.* **Rationale**: Participating in human-centric B2C data ecosystems inevitably requires the data subjects to extensively engage with the system. Due to the multitude and complexity of manual interactions, one can assume that humans' processing capacities, convenience, and digital competencies will by far be exceeded (Bester et al. 2016). Thus, the provision of usability is decisive in B2C data ecosystems, making automation of PD processing crucial. This implies a *broad consent* of data subjects, begging the question of the system's legal compliance. Taking European data law as our legal lens, consent can only be formulated broadly for situations not requiring a *specific consent*. This entails a hybrid nature of the consent mechanisms that have to be integrated in human-centric B2C data ecosystems. Based on their (use case) specific architectures, one must derive and well-define the particular activity types the data subjects can possibly engage in. These activity types, respectively the system architecture, should be designed to maximize situations allowing for broad consents, while requiring specific consent only for a few dedicated interactions that are crucial for humans' data sovereignty. Regardless of the exact nature of the hybrid consent model, "sufficiently tight" appropriation of broad consents must always be ensured. In principle, human-centric B2C data ecosystems must find an equilibrium between automation and self-determination, while data sovereignty must never be curtailed (e.g., revocable broad consents). Initial suggestions for hybrid consent mechanisms are provided by, e.g., Geller et al. (2022), albeit limited to the clinical context.

DP3: *Provide a data governance framework in B2C data ecosystems that comprises a technically enforceable usage control structure which grants to the data subject ownership-like rights to inserted data, while entailing impermeable system boundaries to support data security.* **Rationale**: Since data subjects must always be in charge of all processing activities related to their data, B2C data ecosystems must enforce usage control technically. This entails the restriction that PD do not pass system boundaries on the data demand side. Technical enforcement of usage control is crucial because data subjects seldom have both the digital literacy and the willingness to trace all their datasets shared with data consumers in the system for checking conducted data processing against mutually agreed usage terms. The integrated usage control must allow data processing while usage restrictions are in place and technically enforced. To this end, B2C data ecosystems require secure runtime and execution environments for PD processing that also shield the actual content data from data consumers' access. In these environments, the processing of PD must always be subject to well-defined processing purposes authorized by the data subjects (i.e., usage policies). To ensure effective processing and utilization of PD, data subjects must be limited by their possibilities to specify such usage policies themselves (i.e., via standardized licenses). Otherwise, the heterogeneity of usage policies attached to various datasets in the ecosystem is very

likely to preclude most data processing activities. In this context, data apps are an already applied method to implement largely standardized processing operations in data ecosystems, e.g., Gaia-X and IDS. However, there are no examples in practice yet for binding them to a limited and pre-defined set of well-defined usage purposes with respect to PD processing (appropriation). Besides the approach to data apps, as leveraged in the RSA, DP3 may also be addressed by methods like *compute-to-data* (i.e., the algorithm/data app is sent to distributed data suppliers who return the data product) and *federated learning* (i.e., machine learning in distributed systems and subsequent accumulation of results).

DP4: *Provide a decentralized infrastructure for B2C data ecosystems that supports system scalability and semantic consistency of data and services, thus ensuring their interoperability, processability, and portability in the ecosystem.* **Rationale**: Human-centric B2C data ecosystems should store PD in the sphere of the data subjects as decentral system endpoints. To this end, federated infrastructures, that are open, distributed, and shared, provide adequate means for technical implementation. By their design, such infrastructures should favour data security, privacy, and trustworthiness (Nagel et al. 2021), while supporting findability, interoperability, and portability of both services and data in the ecosystem. This implies the utilization of common and controlled vocabularies for data and service descriptions to ensure semantic consistency and a shared understanding among actors. Following applied practices in B2B contexts (Otto et al. 2019), B2C data ecosystems may reuse existing vocabularies and standards where possible to foster the actors' acceptance. Actors relying on the same vocabularies to structure and formulate both data and service descriptions significantly enhance the functionality and operativeness of the ecosystem. Ultimately, a decentralized infrastructure which emphasizes scalability and semantic consistence favours a level playing field for sovereignly sharing, monetizing, and utilizing PD in human-centric B2C data ecosystems. Best practices like SOLID, IDS, and Gaia-X provide orientation. The DP avoids an ecosystem with a few centralized data providers and thus the concentration of data.

DP5: *Provide trust mechanisms in B2C data ecosystems that ensure transparency of data processing and rigor minimization of information asymmetries to guarantee fairness and confidence among actors as well as traceability of data processing.* **Rationale**: Trust arises as a critical factor in all kinds of ecosystems (Otto et al. 2019) but is particularly relevant in B2C contexts (Rantanen and Koskinen 2020). To achieve trust of the data suppliers, a B2C data ecosystem needs mechanisms for making transparent data processing and consent information, e.g., for tracking processing purposes and consent records. Recorded proofs must ensure the highest levels of comprehensiveness to actually provide added value. Their required granularity depends on the use case and should be determined under the aegis of legal and ethical experts. The comprehensiveness of records can be facilitated by using, for instance, graphical and intuitive user interfaces, privacy icons, et cetera. To win the trust of data consumers, B2C data ecosystems must rigorously minimize information asymmetries. Data consumers should be provided with as much information as possible, given the restrictions of the data governance framework (see DP3) and implications of the market mechanism (see DP1), e.g., Data Providers and Brokers should rely on anonymous metadata in a data catalogue to disclose central attributes of the Data Owners' content data. Above all, transparency about data quality is of crucial importance for data consumers (Wang and Strong 1996). Thus, a vital mechanism in B2C data ecosystems is services for data quality computation to build the trust of data consumers by providing quality information on single datasets. To compute data quality, *IBM Cloud Pak for Data* and *IBM Information Analyzer* are examples of what such an intermediary service in a B2C data ecosystem might look like.

8 Discussion and Conclusion

By exploring B2C data ecosystems, our research pertains to a novel technology for humanity. We built our RSA in orientation toward federated B2B data ecosystem concepts. This makes sense as these concepts exhibit design implications for B2C equivalents in terms of architectural commonalities which are also applicable in the B2C context. However, since design peculiarities with regard to ethics, law, economics, and technology must be considered in handling PD, we also relied on PDMs and related systems as analysis objects. A B2C data ecosystem must reach the point at which a sufficient number of actors have adopted it and interact with each other. This causes *network effects* which make the system's further rate of adoption self-sustaining (Rogers 1995). In our RSA, the interactions between data suppliers, data demanders, and intermediaries are characterizable as a *two-sided market*. Two-sided markets can be used to extend the concept of network effects by including two or more distinct groups of actors (Parker and Alstyne 2005). We have clearly defined and modelled the interactions of these groups in leveraging the designed B2C data ecosystem to innovate data products not creatable otherwise, while each group can receive an added value through engaging in the data ecosystem (e.g., data products, service fees, or prices for data).

From the DPs' level of abstraction, the RSA conforms with existing B2B data ecosystem initiatives in many

respects (see online appendix I; e.g., Mobility DS, IDS, Catena-X, Resilience and Sustainability DS, ForeSight). Design commonalities are, in particular, the reliance on market mechanisms for data sharing and incentivization (#DP1), decentralized infrastructures, especially for data storage (#DP4), as well as trust mechanisms to ensure transparency of data processing and the minimization of information asymmetries (#DP5). However, the RSA also exhibits significant design discrepancies compared to existing B2B data ecosystem initiatives induced by its explicit focus on integrating participative and data sovereign humans with their PD. It contains novel design knowledge concerning the effective implementation of a hybrid consent solution in the architecture of a human-centric data ecosystem, while emphasizing usability through task automation (#DP2). Furthermore, the RSA's data governance model outlines a feasible approach to technically enforce usage control, building upon standardized processing purposes and impermeable system boundaries. In contrast, B2B data ecosystem initiatives exhibit less restrictive data governance models (i.e., orientation toward access control and broader purposes of data processing) as those are adequate for their objectives and the types of data processed. Compared to the few B2C (e.g., SOLID) and data or stakeholder agnostic approaches to data ecosystems (e.g., Gaia-X), the RSA accomplishes to narrow down a set of abstract design elements which must be implemented by their architectures and (federated) services in this or some similar form, given the integration of humans with their PD.

Our work comprises broad scientific and managerial implications. Both the DPs and the RSA collect knowledge about the design of B2C data ecosystems with a specific emphasis on data sovereignty, which practitioners can use. Significantly, this means that practitioners can reduce the effort they require to design such socio-technical systems based on our work as they can learn from our validated design elements (e.g., Kim (2010)). The DPs are easily understandable prescriptions for actions codifying what we have learned. They state conceptually grounded and empirically validated core principles and aspects of designing human-centric data ecosystems abstractly. The RSA accumulates architectural knowledge, providing practitioners with an interplay of building blocks and prescriptions for their effective design. It also helps practitioners reflect on their approaches and ideate concepts on the one hand and, on the other hand, to create B2C data ecosystems from scratch. It facilitates their development in any domain and with higher technology readiness levels (TRLs). Since our RSA is the first of this class of artifact, it contributes to the accumulation of design knowledge (e.g. vom Brocke et al. (2020)). In principle, all B2C environments relying on PD sharing to innovate data-driven products could benefit from our RSA and the reflective DPs, e.g., intermodal traveling, precision medicine, or personalized finance. Conclusively, we provide initial design knowledge for transferring the B2B data ecosystem paradigm to the complex and restricted B2C periphery. We offer inductive insights and a deep understanding of implications which might arise when integrating data-sovereign humans into data ecosystems, thus establishing a solid basis for future research. We see this as a valuable contribution to the scientific community against the background of large-scale European initiatives (e.g., Gaia-X, Catena-X), spurring a plethora of novel research and industry projects that can profit from our work. As a result, we see a potential for future research that can use our findings while refining and extending them. Additionally, data sovereignty issues, especially with PD, are a core object of interest in (European) legislation, which we underline with validated knowledge on how to design these types of data ecosystems effectively.

The study is subject to the following limitations. Firstly, we applied design recovery on samples of PDMs and data ecosystems which might not be representative and are surely not conclusive. Thus, we could have missed functionalities important for B2C data ecosystems. Currently, the *IDSA Data Space Radar* lists between five to ten active data ecosystems, while the number of operative PDMs is unknown. However, precise information about both groups of analysis objects is difficult to obtain. Therefore, we emphasized data ecosystems that assume a standardization character and PDMs that either operate under European jurisdiction or explicitly postulate data sovereignty as a guiding concept. Secondly, as our research area comprises an interdisciplinary and wide field of topics, we must assume to have analyzed a fraction of literature that could be associated only. Thirdly, our research builds upon the assumption that the currently emerging generation of PDMs remains viable. The falsification of this assumption might negatively affect the validity of our artifact. Moreover, our study entails subjectivity issues particularly in terms of the derivation of design implications from literature, the extraction of design features from analysis objects, and the coding of interviews. Lastly, a complete technical proof of concept for the artifact has not been provided, begging the question of its technical feasibility. Yet, this proof is to a certain degree ensured by the applied methods of design recovery and prototyping.

Future research should continue to primarily focus on DSR studies by developing models and technical instantiations with increasing TRLs to accumulate more and concretize existing design knowledge about (human-centric) B2C data ecosystems. By now, our RSA and the reflective DPs represent sets of initial design hypotheses that require more validation and extension. Besides that,

supplementary data sources (e.g., interviews, case studies, literature analyses) should be used in additional domains, to triangulate a more comprehensive look into the topic. This encompasses field tests (e.g., usability testing, business model analyses) to evaluate and refine our proposed RSA and the reflective DPs. Ultimately, the aptitude of human-centric B2C data ecosystems to enhance the current state of PD utilization in our data economy is yet to examine. We would like to emphasize that this disruptive concept can substantially contribute to a fair, trustworthy, and liberal data economy with sovereign humans, which indicates its merit as a technology for humanity.

Funding Open Access funding enabled and organized by Projekt DEAL

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12599-023-00816-9.

References

Abu-Matar M (2016) Towards a software defined reference architecture for smart city ecosystems. In: 2016 IEEE International Smart Cities Conference, Trento

Aseri DAM (2020) The implication of the European Union's general data protection regulation on the global data privacy. J Theor Appl Inf Technol 98(4):692-702

Asif N (2003) Reverse engineering methodology to recover the design artifacts: a case study. Softw Eng Res Pract 2:932–938

Bataineh AS, Mizouni R, Bentahar J, El Barachi M (2020) Toward monetizing personal data: a two-sided market analysis. Fut Gen Comput Syst 111:435–459

Bester J, Cole CM, Kodish E (2016) The limits of informed consent for an overwhelmed patient: clinicians' role in protecting patients and preventing overwhelm. AMA J Ethics 18(9):869–886

Bhattacherjee A (2012) Social science research. Principles, methods, and practices. Scholar Commons, Open Textbook Library, University of South Florida, Tampa

Biggerstaff TJ (1989) Design recovery for maintenance and reuse. Comput 22(7):36–49

Birch K, Cochrane DT, Ward C (2021) Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech. Big Data Soc 8(1):20539517211017308

Chandra L, Seidel S, Gregor S (2015) Prescriptive knowledge in IS research: conceptualizing design principles in terms of materiality, action, and boundary conditions. In: Proceedings of 48th Hawaii International Conference on System Sciences, Kauai, pp 4039–4048

Chikofsky EJ, Cross JH (1990) Reverse engineering and design recovery: a taxonomy. IEEE Softw 7(1):13–17

Cloutier R, Muller G, Verma D, Nilchiani R, Hole E, Bone M (2009) The concept of reference architectures. Syst Eng 14(3):14–27

Cohendet P, Meyer-Krahmer F (2001) The theoretical and policy implications of knowledge codification. Res Policy 30(9):1563–1591

Currie WS (2011) Units of nature or processes across scales? The ecosystem concept at age 75. New Phytol 190(1):21–34

Geller S, Müller S, Scheider S, Woopen C, Meister S (2022) Value-based consent model: a design thinking approach for enabling informed consent in medical data research. In: Proceedings of the 15th international joint conference on biomedical engineering systems and technologies, pp 81–92. https://www.scitepress.org/Papers/2022/108280/108280.pdf

Governatori G, Rotolo A, Villata S, Gandon F (2013) One license to compose them all: a deontic logic approach to data licensing on the web of data. In: International semantic web conference (8218), Sydney, pp 151–166

Gregor S, Jones D (2007) The anatomy of a design theory. J Assoc Inf Syst. https://doi.org/10.17705/1jais.00129

Gregor S (2009) Building theory in the sciences of the artificial. In: Proceedings of the 4th international conference on design science research in information systems and technology, Philadelphia, Article 4

Hanssen GK, Dyba T (2012) Theoretical foundations of software ecosystems. In: Proceedings of the international workshop on software ecosystems, Cambridge, pp 6–17

Hennink MM, Kaiser BN, Marconi VC (2017) Code saturation versus meaning saturation: How many interviews are enough? Qual Health Res 27(4):591–608

Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Q 28(1):75–105. https://doi.org/10.2307/25148625

Hevner AR (2007) A three cycle view of design science research. Scand J Inf Syst 19(2):87-92

Hummel P, Braun M, Dabrock P (2021) Own data? Ethical reflections on data ownership. Philos Technol 34(3):545–572

Iivari J (2007) A paradigmatic analysis of information systems as a design science. Scand J Inf Syst 19(2):39–64

Jacobides MG, Cennamo C, Gawer A (2018) Towards a theory of ecosystems. Strateg Manag J 39(8):2255–2276

Jansen S, Cusumano MA, Brinkkemper S (eds) (2013) Software ecosystems: analyzing and managing business networks in the software industry. Elgar, Cheltonham

Kim H (2010) Effective organization of design guidelines reflecting designer's design strategies. Int J Ind Ergon 40(6):669–688

Kortum H, Kohl T, Hubertus D, Hinz O, Thomas O (2022) A platform framework for the adoption and operation of ML-based smart services in the data ecosystem of smart living. In: Demmler D et al (eds) Informatik. Gesellschaft für Informatik, Bonn, pp 361–377

Koskinen J, Knaapi-Junnila S, Rantanen MM (2019) What if we had fair, people-centred data economy ecosystems? In: 2019 IEEE SmartWorld, IEEE, pp 329–334

Kruchten P (1995) Architectural blueprints–the "4+1" view model of software architecture. IEEE Softw 12(6):42–50

Lauf F, Scheider S, Bartsch J, Herrmann P, Radic M, Rebbert M, Nemat AT, Schlueter-Langdon C, Konrad R, Sunyaev A, Meister S (2022) Linking data sovereignty and data economy: arising areas of tension. In: Proceedings of the 17th international

conference on wirtschaftsinformatik. https://aisel.aisnet.org/wi2022/it_for_development/it_for_development/19
Lee D (2014) Building an open data ecosystem: an Irish experience. In: Proceedings of the 8th international conference on theory and practice of electronic governance, Guimaraes, pp 351–360
Leidner DE, Tona O (2021) The CARE theory of dignity amid personal data digitalization. MIS Q 45(1):343–370
Li W, Badr Y, Biennier F (2012) Digital ecosystems: challenges and prospects. In: Proceedings of the international conference on management of emergent digital ecosystems, Addis Ababa, pp 117–122
Lindman J, Kinnari T, Rossi M (2015) Business roles in the emerging open-data ecosystem. IEEE Softw 33(5):54–59
March ST, Smith GF (1995) Design and natural science research on information technology. Decis Support Syst 15(4):251–266. https://doi.org/10.1016/0167-9236(94)00041-2
March ST, Storey VC (2008) Design science in the information systems discipline: an introduction to the special issue on design science research. MIS Q 32(4):725–730
Maritain J (1950) The concept of sovereignty. Am Political Sci Rev 44(2):343–357
Meister S, Otto B (2019) Digital life journey: a framework for a self-determined life of citizens in an increasingly digitized world. Basic Research Paper. https://www.digitallifejourney.de/en/
Meth H, Mueller B, Maedche A (2015) Designing a requirement mining system. J Assoc Inf Syst 16(9):2
Metzger A (2020) A market model for personal data: state of play under the new directive on digital content and digital services. In: Lohsse S et al (eds) Data as counter-performance – contract law 2.0? Nomos
Moiso C, Minerva R (2012) Towards a user-centric personal data ecosystem: the role of the bank of individuals' data. In: 16th International Conference on Intelligence in Next Generation Networks, Berlin, pp 202–209. https://doi.org/10.1109/icin.2012.6376027
Möller F, Guggenberger TM, Otto B (2020) Towards a method for design principle development in information systems. In: Proceedings of the 15th international conference on design science research in information systems and technology, Kristiansand, pp 208–220. https://link.springer.com/chapter/10.1007/978-3-030-64823-7_20
Nagel L, Lycklama D, Ahle U (2021) Design principles for data spaces. Position Paper, https://design-principles-for-data-spaces.org/. Accessed 4 May 2022
Oehler A (2016) Chancen der selbstbestimmten Datennutzung. Wirtschaftsdienst 96(11):830–832
Oliveira M, Lima G, Lóscio BF (2019) Investigations into data ecosystems: a systematic mapping study. Knowl Inf Syst 61(2):589–630
Otto B, Lis D, Jürjens J, Cirullies J, Opriel S, Howar F, Meister S, Spiekermann M, Pettenpohl H, Möller F (2019) Data ecosystems. Conceptual foundations, constituents case studies and recommendations for action. White Paper, Fraunhofer ISST. https://www.isst.fraunhofer.de/content/dam/isst-neu/documents/Publikationen/StudienundWhitePaper/FhG-ISST_DATA-ECOSYSTEMS.pdf
Parker GG, Van Alstyne MW (2005) Two-sided network effects: a theory of information product design. Manag Sci 51(10):1494–1504
Parra-Arnau J (2018) Optimized, direct sale of privacy in personal data marketplaces. Inf Sci 424:354–384. https://doi.org/10.1016/j.ins.2017.10.009
Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S (2007) A design science research methodology for information systems research. J Manag Inf Syst 24(3):45–77. https://doi.org/10.2753/MIS0742-1222240302
Ploug T, Holm S (2016) Meta consent-a flexible solution to the problem of secondary use of health data. Bioethics 30(9):721–732
Pratt MG (2008) Fitting oval pegs into round holes. Organ Res Meth 11(3):481–509
Rantanen M, Koskinen J (2020) Respecting the individuals of data economy ecosystems. In: Proceedings of the 8th international conference on well-being in the information society, Turku, pp 185–196
Rantanen M, Hyrynsalmi S, Hyrynsalmi SM (2019) Towards ethical data ecosystems: a literature study. In: IEEE international conference on engineering, technology and innovation, Valbonne Sophia-Antipolis
Reidt A (2018) Referenzarchitektur eines integrierten Informationssystems zur Unterstützung der Instandhaltung. Dissertation, Technische Universität München
Rogers EM (1995) Diffusion of innovations. Free Press, New York
Sambra AV, Mansour E, Hawke S, Zereba M, Greco N, Ghanem A, Zagidulin D, Aboulnaga A, Berners-Lee T (2016) Solid: a platform for decentralized social applications based on linked data. Tech. Rep, MIT CSAIL & Qatar Computing Research Institute
Scheider S, Lauf F, Geller S (2023) Data sovereign humans and the information economy: towards design principles for human centric B2C data ecosystems. In: Proceedings of the 56th Hawaii international conference on system sciences, pp 3725–3734. https://scholarspace.manoa.hawaii.edu/items/bae9367e-2959-48d9-8e22-06d714d0bfcd
Sonnenberg C, vom Brocke J (2012) Evaluations in the science of the artificial-reconsidering the build-evaluate pattern in design science research. In: Proceedings of the 7th international conference on design science research in information systems, Las Vegas, pp 381–397
Spiekermann S (2016) Ethical IT innovation-a value-based system design approach. CRC Press, New York
Spiekermann S, Acquisti A, Böhme R, Hui K-L (2015) The challenges of personal data markets and privacy. Electron Mark 25(2):161–167
Strauss A, Corbin JM (1990) Basics of qualitative research: grounded theory procedures and techniques. Sage, Newbury Park
vom Brocke J, Winter R, Hevner A, Maedche A (2020) Accumulation and evolution of design knowledge in design science research: a journey through time and space. J Assoc Inf Syst 21(3):520–544
Wang P (2021) Connecting the parts with the whole: toward an information ecology theory of digital innovation ecosystems. MIS Q 45(1):397–422
Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 12(4):5–33