Interoperable Open Data Platforms: A Prototype for Sharing CKAN Data Sources

Sebastian Becker[a] and Marcel Altendeitering[b]
Fraunhofer ISST, Speicherstraße 6, Dortmund, Germany

Keywords: Data Platform, Data Sharing, Data Space, Data Management.

Abstract: Open data promotes transparency, accountability, and innovation in organizations and represents a central element of modern data management, supporting informed decision-making. CKAN is the world's leading open-source data portal, widely used on national and local open-data platforms. However, CKAN installations are usually operated insularly, with limited interoperability and interaction between multiple instances. The induced separation is based on incompatible data models, leading to complex searches that include several open data platforms. In this paper, we describe a prototype that realizes an interoperability layer between CKAN instances and connects them in a data space. To create the intended solution, we relied on the Eclipse Dataspace Components (EDC) open-source project and present details on our architectural approach and implementation. For evaluation, we conducted a series of focus group discussions with stakeholders of our prototype. We received mostly positive feedback on our developments, and the participants agreed that our solution could lead to an improved interoperability of open-data platforms.

1 INTRODUCTION

In today's digital age, the accessibility and utilization of data sources have become increasingly important. Leveraging a large, high-quality data basis is vital for organizations as data fuels many innovations, is required for automated decision-making (e.g., ML and AI), and can secure organizational agility (Amadori et al., 2020; Gröger, 2021; Kabalisa and Altmann, 2021). Data ecosystems and data spaces are essential concepts for breaking data silos and enabling data reuse for mutual benefit. Realizing such inter-organizational data flows can help to innovate cooperatively and exploit new business opportunities (S. Oliveira et al., 2019; Azkan et al., 2020).

However, the concepts of data sharing and interoperability still need to be introduced to open data platforms such as CKAN, which often reside insular and disconnected from other open data platforms. This leads to complex data access and limited data availability to data consumers. A technical solution that makes open data sources available to other data platforms is currently missing as most solutions strongly focus on a single domain (for Economic Affairs and Action, 2024).

Our study aims to address this research gap and contribute to the design and realization of interoperable open data platforms. Consequently, our research question reads as follows:

Research Question: *How to design and realize interoperable open data platforms?*

In response to the proposed research question, we conceptually developed and prototypically implemented a software solution that enables the sharing of CKAN-based data sources between different instances. For this purpose, we applied the architectural concept of data spaces to an exemplary CKAN instance. Specifically, we extended the plug-in architecture of the Eclipse Dataspace Components (EDC) (Foundation, 2024a) project with a custom extension that integrates with CKAN (CKAN, 2024b) and follows adequate information models. This way, our developments can lead to larger data bases that can facilitate the creation of data-intensive applications and services. We used CKAN and the EDC as a base for our developments as they are both open-source software projects and widely established in their respective fields.

We evaluated our prototype qualitatively through a

[a] https://orcid.org/0009-0001-6309-7351
[b] https://orcid.org/0000-0003-1827-5312

Becker, S. and Altendeitering, M.
Interoperable Open Data Platforms: A Prototype for Sharing CKAN Data Sources.
DOI: 10.5220/0012812100003756
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 486-493
ISBN: 978-989-758-707-8; ISSN: 2184-285X
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

series of focus group discussions, including ten stakeholders involved in the development of our solution (Krueger and Casey, 2014). These discussions helped us continuously improve the application and ensure our design and development is in line with the desired goal. Overall, our results are well-received, and we have successfully developed a suitable prototype. We also identified multiple areas for improvement that can inform future work.

The remainder of this paper is structured as follows: First, in section 2, we present the background of our study, describing open data platforms, the concepts for data sharing and data basis, and, specifically, the EDC project we relied on. In section 3, we describe our prototype in terms of the architectural approach and implementation details. We present and discuss the results of our qualitative evaluation in section 4. Finally, section 5 highlights the contributions of our study and its limitations and shows areas for future work.

2 BACKGROUND

2.1 Open Data Platforms

In recent times, open data has become increasingly more prevalent as it can be used in a number of different fields and has a high potential of enabling economic growth (Smith et al., 2016). Apart from economic aspects it is also often times used by governments in order to increase transparency and therefore trust. (Bertot et al., 2010). Governments can publish a variety of different data sets, such as information about the general population, receipts and expenditures, education, environment, science and many more. Germany for example currently hosts 97.125 data sets (Germany, 2024). There is also a European data collection, currently holding 1.709.067 data sets from 35 different countries across Europe as well as a few non-European or European Union data sets (Union, 2024a).

In addition to government or financial data, another important sector for open data is public service e.g. public transportation, since it can help to improve data flow between different parties and raise the level of transparency (Smith et al., 2016).

Having lots of different application possibilities for open data also calls for open data platforms, that are capable of displaying available data sets and allow individuals to search, filter and explore the available data. An open data platform should be able to accomplish those tasks and can additionally supply an API, that allows automated access to data for software applications (Union, 2024b).

The world's leading open source data management system, that fulfills those criteria is called CKAN (CKAN, 2024a). This open data platform is used by the governments of the United States, Canada and Australia as well as other big corporations and governments. Therefore, our work focuses on importing catalogs from a CKAN instance into a data space in order to gain access to a widely used standard as well as interconnecting multiple CKAN instances to be able to form a larger collection of data sets.

2.2 Data Sharing & Data Spaces

Extensive, exchangeable, and high-quality data sets are crucial for data-intensive applications and are valuable assets for organizations (Gröger, 2021; Otto et al., 2019). External data sets are often needed to complement internal data sets and drive data-intensive applications and innovation (Gröger, 2021; Hosseinzadeh et al., 2020). For example, training large language models (LLMs) requires substantial data that most organizations cannot acquire independently. Therefore, "data exchange between [and within] companies is an essential feature of digitization and data economy," (Hosseinzadeh et al., 2020) (p.1), always facing challenges related to data management, quality, and governance (Gröger, 2021; Janev et al., 2021). The main issue here is that data owners are often reluctant to share their data, even if there are contracts and agreements in place to dictate its usage. As a result, data providers are constantly "afraid of losing control" (Chakrabarti et al., 2018) (p.1) over how their data is used. Ensuring the interests of the data provider from a technical standpoint and building trust among participants in a data ecosystem is particularly challenging.

To address these aspects, data space initiatives, such as the International Data Spaces Association (IDSA) emerged and promised to enable the sovereign sharing of data within a data ecosystem. Realizing this promise required the definition of a shared protocol, which ensures a common understanding of data sharing and interoperability between the participants of a data space. For this purpose, the IDSA created the so-called data space protocol (IDSA, 2024b). The data space protocol is "a set of specifications designed to facilitate interoperable data sharing between entities governed by usage control and based on Web technologies" (IDSA, 2024b) (p.1). Specifically, it specifies how data sets are described using standards such as the Data Catalog Vocabulary (DCAT) and the Open Data Rights Language (ODRL) (Bader et al., 2020). Furthermore, it specifies how data exchange agreements are expressed and negotiated and how data sets are accessed using transfer protocols.

There are multiple solutions available that implement the IDSA data space protocol (IDSA, 2024a). One of the projects are the Eclipse Dataspace Components (EDC). We decided to use the EDC for our prototype as the EDC-project is open-source, has an active community, and is widely established in industrial and academic projects (e.g., (Catena-X, 2024)).

2.3 Eclipse Dataspace Components (EDC)

The Eclipse Dataspace Components (EDC) (Foundation, 2024a) is a framework maintained and developed by the Eclipse Foundation as well as a number of big companies and research institutes, such as Microsoft, SAP and Fraunhofer. The framework offers an implementation of the IDSA data space protocol and is setup in a way that offers developers to build their own and highly customizable data space by including only the necessary components. In addition to the high flexibility of choosing the fitting components it is also easily extendible with new extensions.

One of the key principles of the EDC framework is the separation of Control Plane and Data Plane, which serves the purpose of reducing overhead and clearly separating the communication between participants into two channels. The Data Plane is solely used for sharing data in the end, after the organizational side is fulfilled, while the Control Plane is used to arrange the data exchange by negotiating contracts, that are associated with the assets that should be shared.

Another important aspect of this framework is the implementation of a federated catalog, which is a two-folded system consisting of a cache and a crawler. Each participant of the data space has their own federated catalog which crawls the catalog of the other participants on a regular basis in order to fill the cache with the information of available data sets where the attached policies are fulfilled.

The idea of using a data space and a unified communication protocol is especially important when talking about a decentralized architecture, where individual parties that are involved communicate between each other. A minimalistic implementation of a data space using the EDC framework only requires very little central components such as a registration service, that is able to register users as well as a DID-server, that provides them with so-called Decentralized Identifiers (DID).

Such a minimalistic approach to a data space is implemented in form of the Minimum Viable Dataspace (MVD), which can be seen in Figure 1. The data space with its Data Plane and Control Plane consists of the Registration Service, a DID-server and multiple participants, who each have a federated catalog and UI.

3 PROTOTYPE

As described in Section 2.3 the EDC offer a framework which allows the implementation of a highly customizable data space. This project was aimed at facilitating a demonstrator in order to showcase the capabilities of mirroring a CKAN catalog into a data space. Therefore, we used the Minimum Viable Dataspace (MVD) (Foundation, 2024c) as a base since it already implements a registration service and a DID-server, which are used for registering new participants in the data space and managing the exchange and supply of DID documents to the user. Furthermore, the MVD implements a rudimentary graphical interface in form of a dashboard which is capable of showing available policies, assets, contract definitions and contract negotiations and is a helpful tool for understanding the underlying principles of the data exchange inside of the data space.

3.1 Architecture

Our approach of integrating a CKAN file catalog and our CKAN importer extension into the Minimum Viable Dataspace can be seen in Figure 2. For demonstrating purposes two participants are connected via a data space each of them using an implementation of the EDC framework. When first starting up the data space, the participants are registered as users by using the registration service of the MVD. When the users are registered, each of them is creating their own instance of a federated catalog which holds the metadata of the owned data sets and data sets that are made available by other participants where the contract definitions allow it.

Usually, the catalog fetches the data set information from some kind of data sink. In our case though the data needs to be harvested from the CKAN catalog by our importer extension which is done by requesting the appropriate catalog endpoint of the local CKAN instance. The requested catalog holds the information of all locally available data sets of this participant. In addition to actual content of the data sets, a field that specifies the policy, under which the data set is allowed to be shared with other participants of the data space, is supplied as well. In the example scenario *Data set 3* has a different policy, that prohibits the ex-

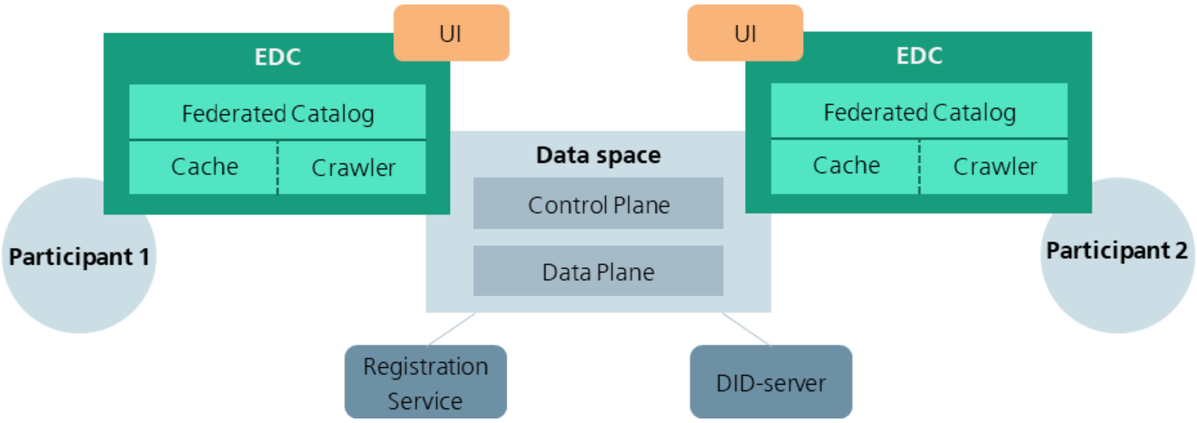

Figure 1: Minimum Viable Dataspace (MVD).

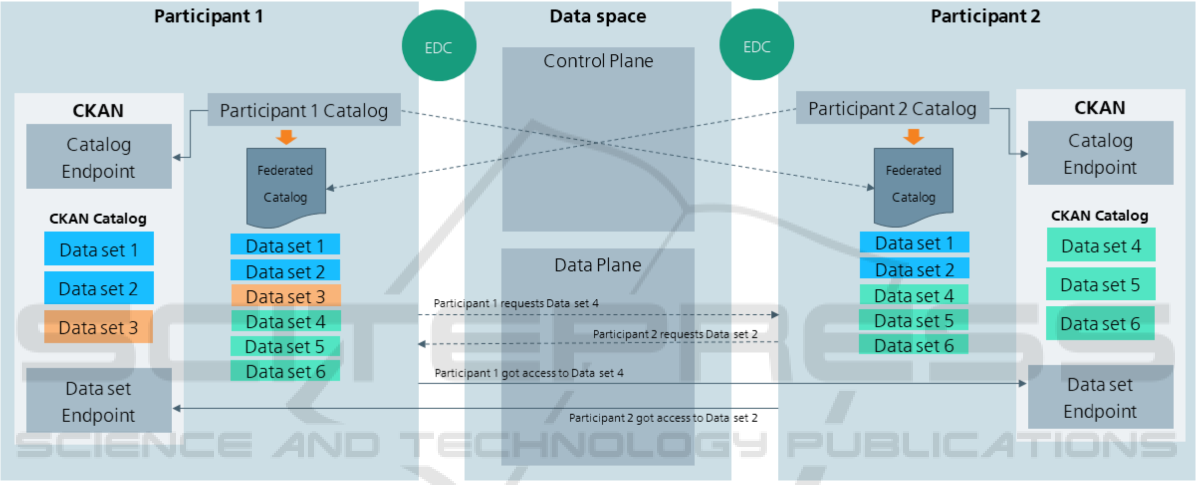

Figure 2: Filling the federated catalog with data from a CKAN catalog.

change of this particular data set with *Participant 2*, while all other data sets use some kind of policy that allows the other user to use this data set. Therefore, the federated catalog from *Participant 1* contains six different data sets, while the other federated catalog contains only five.

After the initial setup, each of the registered users has a federated catalog which is filled with metadata information about all available data sets in their own CKAN instance as well as with the information about data sets from other participants' CKAN instances where the restrictions made by the attached policies are respected.

In our example, *Participant 1* can now request *Data set 4* since it is made available by *Participant 2*. In order to do this, the contract negotiation phase, which is still part of the control plane, is started. If both parties agree on the contract, which includes to respect the usage policies attached to this data set, the data plane is used to actually transfer the data. The CKAN instance also contains a data set endpoint, which is part of the metadata information given by the data set. If the contract negotiation was successful, the data set is provided to the requesting user by leveraging the HTTP Data Plane (Foundation, 2024b). By using this extension, the data provider fetches the data by requesting the data set endpoint and forwarding this information to the data plane which in turn transfers it to the consumer. Using the HTTP Data Plane is an important step, as only the provider of a data set has the access rights to the locally hosted CKAN instance.

3.2 Implementation

The proposed architecture heavily builds upon filling the federated catalog using the CKAN importer extension, which is a structured three-phase process. The CKAN instance provides its catalog in multiple different formats (CKAN, 2024b), giving us the oppor-

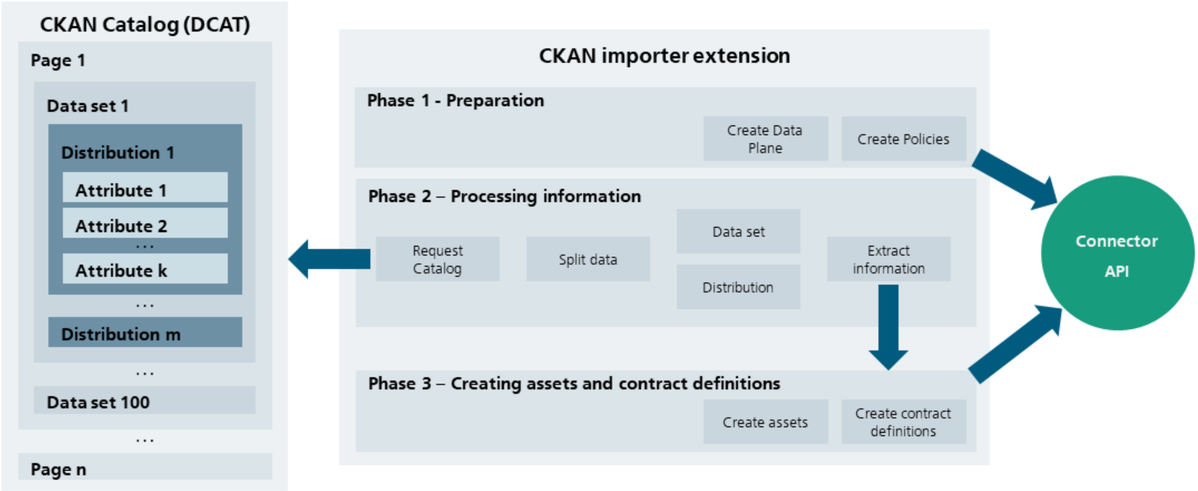

Figure 3: DCAT catalog structure (*left*) and 3-step implementation of the CKAN importer extension (*right*).

tunity to use the DCAT format by requesting the catalog endpoint and supplying the .ttl ending.

The structure of a catalog in the DCAT format as well as the three-phase approach of importing this data into the data space can be seen in Figure 3.

DCAT Structure. A catalog that is using the DCAT format has a layer-like structure, which can be compared to a folder structure on computers. The whole catalog contains n pages, where n is controlled by the amount of data sets, since every page is able to hold up to 100 data sets. Therefore, requesting the catalog might require multiple requests in order to fetch different pages and retrieve all available data sets. Every data set is made up of at least one distribution but can contain m distributions, where m is not limited. In addition to the distributions, each data set contains some more metadata like the creator of the data set, description etc. Every distribution on the other hand is made up of k attributes, where k is not limited again. In our case, the *accessUrl* is the most important attribute, as this information is necessary for the previously described HTTP Data Plane to find the corresponding resource on the provider side. Apart from this attribute, other important ones include the format of the file, the title and the modified timestamp. The latter is also used for comparison with already existing assets in the data space in order to avoid unnecessary API calls.

Phase 1: Preparation. The first phase in the proposed process is the preparation phase, which lays the foundation for subsequent activities. A data plane is created, which will later be used in order to share data between the participants. Furthermore, a previously defined set of policies is created. This pool of policies matches the policies that are attached to the data sets in the CKAN catalog. The here created policies can be attached to assets in phase 3 in order to build the contract definitions for every asset. Both processes are done by sending POST requests to the Connector API.

Phase 2: Processing Information. The second phase involves processing the information gathered by requesting all pages of the local CKAN instance. After requesting, the catalog data is split into data sets and distributions and is then processed individually. The available information is extracted and transformed into JSON objects that can be used in phase 3 for the asset creation process.

Phase 3: Creating Assets and Contract Definitions. The final step focuses on actually creating the assets in the data space and adding the corresponding policies in order to create contract definitions for every data set. Both processes make use of the Connector API again by calling different endpoints. The creation of the assets makes use of the JSON objects, which are the product of phase 2. Before creating an any asset though, all already available assets in the data space are requested using the Connector API in order to check if the asset exists already, which can be achieved by comparing the ids. If the new id does not exist yet, the Connector API is used to create it. Otherwise, the modified timestamp given with the metadata of each data set or distribution is compared to the already existing asset in the data space. If the timestamp matches, no further action is necessary. Otherwise, the asset is updated by using the previously extracted information. After successfully creating or updating an asset, the by the data set supplied policy is added to the asset in order to create a contract definition. Only after this step the asset is viewable for other participants in the data space that fulfill the criteria of the added policy.

4 EVALUATION

Our evaluation approach follows the guidelines of (Kitchenham et al., 2002) and (Kitchenham et al., 2004). In the studies, the authors formulate guidelines for conducting experimental and evidence-based software engineering research that aims to critically assess the validity, impact, and applicability of prototypical software solutions. Following (Kitchenham et al., 2002) (p.734), a common problem of software engineering research is "... collecting the experimental outcome measures". Using a well-defined and structured data collection process is important to secure the validity and traceability of the results.

Our evaluation approach follows the guidelines of (Kitchenham et al., 2002) and (Kitchenham et al., 2004). In the studies, the authors formulate guidelines for conducting experimental and evidence-based software engineering research that aims to critically assess the validity, impact, and applicability of prototypical software solutions. Following (Kitchenham et al., 2002) (p.734), a common problem of software engineering research is "... collecting the experimental outcome measures". Using a well-defined and structured data collection process is important to secure the validity and traceability of the results.

To gather empirical feedback on our prototype and assess its validity and applicability in a real-life context, we conducted regular focus group discussions. Focus-groups are well-suited for collecting empirical feedback as they have a high-degree of external validity and can replicate discussions participants have in their daily lives (Krueger and Casey, 2014; Hollander, 2004). Over a course of three months, we conducted weekly focus group discussions, including a presentation of the prototype and subsequently a moderated discussion. These meetings consisted of ten participants, representing different stakeholders of our prototype. These included the development team, product owner, requirements engineers, data engineers, and managerial stakeholders. After each discussion, we prioritized the feedback we obtained and extended the prototype accordingly. The following two subsections summarize the positive and negative feedback we received in the final focus group discussion at the end of the project.

Overall, the feedback we received was mostly positive. Most importantly, we successfully implemented a prototype that allows to share CKAN data sources with external data consumers and, hence, achieved our research goal. All participants positively highlighted this possibility and emphasized the potentials of combining open data platforms and data spaces. Considerably, the participants liked the possibility to create policies for data sources that can limit data sharing to specific geographic areas or time frames.

However, the participants also mentioned several aspects that require further attention and should be extended in the future. (1) The participants noted that an individual user management and authorization would be beneficial for integrating the prototype in an established system landscape. (2) The data usage policies are currently created through an API. The participants envisioned a possibility to create the usage policies graphically using a suitable user interface. (3) Search functionalities within the data space are currently limited to the metadata of data sets. The participants suggested to conduct a complete search, including the data sets, mentioning this would help to create a more useful solution. The feedback helps us create a more advanced prototype as part of future work (see also section 5.2).

5 CONCLUSION

5.1 Contributions

We have successfully developed a prototype that facilitates the sharing of CKAN data sets, enabling the creation of inter-organizational CKAN-based data hubs accessible to external users and data consumers. Our prototype was evaluated qualitatively in a focus group discussion and received largely positive feedback.

From a *scientific* perspective, we have conceptually designed and prototypically implemented an artifact (i.e., an EDC extension) that addresses the need for improved interoperability in open-data platforms. By integrating the concepts of established open-source data platforms and interoperable data spaces, we present an empirical and evaluated software artifact. By highlighting the feedback we gained during evaluation, we offer opportunities for future work, leading to further research attention on the areas of open data and data spaces.

Additionally, we offer *managers* architectural concepts, implementation details, and insights into our course of action for realizing our prototype. Practitioners can use this knowledge to promote sharing their own CKAN data sets and opening existing data hubs to external data consumers. Specifically, this knowledge can assist data management system operators in expanding their data and system architectures to align with the concepts of data ecosystems and data spaces. Consequently, this can lead to more extensive data bases and support the creation of applications based upon them.

5.2 Limitations & Future Work

Despite following a rigorous approach, our study is subject to several limitations, which offer paths for future work. Most importantly, our developments are shaped by the context in which we developed our solution and thus limited in this regard. Implementing our prototype in further scenarios can help us identify further requirements and lead to more profound design knowledge. Additionally, we plan to extend our prototype functionally. These developments can include: (1) a more in-depth search functionality that includes all relevant metadata, (2) allowing to filter data sets based on file size, file type, last modified, etc., (3) conducting tests on high-volume data sets and runtime analyses.

Concerning our methodological approach, our evaluation results are based on qualitative feedback from participants working in the organizational context of our prototype. Additional studies featuring a more diverse group of participants can help us gather additional feedback and improve the interoperability of our approach. Based on such insights coming from multiple cases, we plan to formulate a robust set of design principles that can assist other researchers and practitioners in designing and developing data sharing solutions for CKAN-based data platforms.

REFERENCES

Amadori, A., Altendeitering, M., and Otto, B. (2020). Challenges of data management in industry 4.0: A single case study of the material retrieval process. In *Business Information Systems: 23rd International Conference, BIS 2020, Colorado Springs, CO, USA, June 8–10, 2020, Proceedings 23*, pages 379–390. Springer.

Azkan, C., Möller, F., Meisel, L., and Otto, B. (2020). Service dominant logic perspective on data ecosystems-a case study based morphology. In *Proceedings of the 28th European Conference on Information Systems*.

Bader, S. R., Pullmann, J., Mader, C., Tramp, S., Quix, C., Müller, A. W., Akyürek, H., Böckmann, M., Imbusch, B. T., Lipp, J., Geisler, S., and Lange, C. (2020). The International Data Spaces Information Model – An Ontology for Sovereign Exchange of Digital Content. In *International Semantic Web Conference*, pages 176–192. Springer.

Bertot, J. C., Jaeger, P. T., and Grimes, J. M. (2010). Using icts to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*, 27(3):264–271.

Catena-X (2024). Catena-X. (Accessed: 03.04.2024).

Chakrabarti, A., Quix, C., Geisler, S., Pullmann, J., Khromov, A., and Jarke, M. (2018). Goal-Oriented Modelling of Relations and Dependencies in Data Marketplaces. In *Proceedings of the 11th Inter-national Workshop i* co-located with the 30th International Conference on Advanced Information Systems Engineering*.

CKAN (2024a). CKAN. (Accessed: 08.04.2024).

CKAN (2024b). CKAN DCAT extension. (Accessed: 08.04.2024).

for Economic Affairs, F. M. and Action, C. (2024). How to share data? Data sharing platforms for orginizations. (Accessed: 08.04.2024).

Foundation, E. (2024a). Eclipse Dataspace Components. (Accessed: 08.04.2024).

Foundation, E. (2024b). HTTP Data Plane. (Accessed: 08.04.2024).

Foundation, E. (2024c). Minimum Viable Dataspace. (Accessed: 05.04.2024).

Germany, G. (2024). GovData Germany. (Accessed: 08.04.2024).

Gröger, C. (2021). There is no ai without data. *Communications of the ACM*, 64(11):98–108.

Hollander, J. A. (2004). The social contexts of focus groups. *Journal of contemporary ethnography*, 33(5):602–637.

Hosseinzadeh, A., Eitel, A., and Jung, C. (2020). A Systematic Approach toward Extracting Technically Enforceable Policies from Data Usage Control Requirements. In *Proceedings of the 6th International Conference on Information Systems Security and Privacy*, pages 397–405.

IDSA (2024a). Data Connector Report. (Accessed: 03.04.2024).

IDSA (2024b). Dataspace Protocol 2024-1. (Accessed: 03.04.2024).

Janev, V., Vidal, M. E., Endris, K., and Pujic, D. (2021). Managing Knowledge in Energy Data Spaces. In *Companion Proceedings of the Web Conference 2021*, pages 7–15, New York, NY, USA. ACM.

Kabalisa, R. and Altmann, J. (2021). Ai technologies and motives for ai adoption by countries and firms: a systematic literature review. In *Economics of Grids, Clouds, Systems, and Services: 18th International Conference, GECON 2021, Virtual Event, September 21–23, 2021, Proceedings 18*, pages 39–51. Springer.

Kitchenham, B. A., Dyba, T., and Jorgensen, M. (2004). Evidence-based software engineering. In *Proceedings. 26th International Conference on Software Engineering*, pages 273–281. IEEE.

Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., and Rosenberg, J. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on software engineering*, 28(8):721–734.

Krueger, R. A. and Casey, M. A. (2014). *Focus Groups: A Practical Guide for Applied Research*. SAGE Publications.

Otto, B., Steinbuss, S., Teuscher, A., and Lohmann, S. (2019). Ids reference architecture model.

S. Oliveira, M. I., Barros Lima, G. d. F., and Farias Lóscio, B. (2019). Investigations into data ecosystems: a systematic mapping study. *Knowledge and Information Systems*, 61:589–630.

Smith, G., Ofe, H. A., and Sandberg, J. (2016). Digital service innovation from open data: Exploring the value proposition of an open data marketplace. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1277–1286.

Union, E. (2024a). European Data. (Accessed: 08.04.2024).

Union, E. (2024b). Open Data Portals. (Accessed: 08.04.2024).