

Article

Collaborative Data Use between Private and Public Stakeholders—A Regional Case Study

Claire Jean-Quartier ^{1,*} , Miguel Rey Mazón ¹ , Mario Lovrić ²  and Sarah Stryeck ^{1,2,*} 

¹ Institute for Data Science and Interactive Systems, Graz University of Technology, 8010 Graz, Austria; m.reymazon@tugraz.at

² Know-Center GmbH, 8010 Graz, Austria; mlovric@know-center.at

* Correspondence: c.jean-quartier@tugraz.at (C.J.-Q.); sarah.stryeck@tugraz.at (S.S.)

Abstract: Research and development are facilitated by sharing knowledge bases, and the innovation process benefits from collaborative efforts that involve the collective utilization of data. Until now, most companies and organizations have produced and collected various types of data, and stored them in data silos that still have to be integrated with one another in order to enable knowledge creation. For this to happen, both public and private actors must adopt a flexible approach to achieve the necessary transition to break data silos and create collaborative data sharing between data producers and users. In this paper, we investigate several factors influencing cooperative data usage and explore the challenges posed by the participation in cross-organizational data ecosystems by performing an interview study among stakeholders from private and public organizations in the context of the project IDE@S, which aims at fostering the cooperation in data science in the Austrian federal state of Styria. We highlight technological and organizational requirements of data infrastructure, expertise, and practises towards collaborative data usage.

Keywords: cooperative data use; public-private data sharing; cross-organizational data exchange; federated resources; collaborative data environment



Citation: Jean-Quartier, C.; Rey Mazón, M.; Lovric, M.; Stryeck, S. Collaborative Data Use between Private and Public Stakeholders—A Regional Case Study. *Data* **2022**, *7*, 20. <https://doi.org/10.3390/data7020020>

Academic Editors: Alexander Kotsev, Marco Minghini, Massimo Craglia, Stefano Nativi and Carlos Granell

Received: 29 October 2021

Accepted: 26 January 2022

Published: 28 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The hidden potential for research and innovation held by data is nowadays widely acknowledged [1,2]. However, collaborative use of data based on data exchange between the parties involved is still far from becoming a standard practice and still faces various challenges [3–7]. In our study, focused on the Austrian Federal State of Styria, we survey regional stakeholders from private and public institutions to identify the factors that facilitate collaborative data use, and investigate the need for data infrastructures, expertise, and training, as well as attitudes toward sharing, which could enable the widespread cooperative use of data in the future.

The project “IDE@S: Innovative Data Environment @ Styria”, funded by the government of the Federal State of Styria, is fostering cooperation between Styrian stakeholders (companies, academia, and public institutions) in data science and collaborative data tools with the goal of creating a distributed data platform that ensures data sovereignty irrespective of how data is exchanged. This is achieved by exploring the current state, existing gaps, and desired development of technical infrastructures, services, human resources (including education and training), as well as governance structure and potential business models. As one of the most innovative regions in the EU, Styria presents a rich mix of higher education and research institutions and industrial clusters (in sectors like e.g., automotive, green tech, health, or semiconductor technologies), eager to adopt new technologies and innovation strategies, and constitutes the ideal breeding ground for the creation of innovative data ecosystems.

The present study, part of the IDE@S project, provides insight into day-to-day work with data from Styrian stakeholders, investigates their attitude toward data-driven collabor-

oration, and sets the basis to evaluate the need for regional data infrastructures. It also provides an overview of data-related tasks, best practices, and competences among staff. We use the responses to interviews held with Styrian stakeholders in the course of the project to analyze what further developments are required to extend collaboration through data exchange. The paper is organized as follows: After giving background information in Section 2, we introduce the interview framework in Section 3, then show the results of closed and open questions in Section 4, and discuss the main findings in Section 5.

2. Literature Survey

Many actors in the academic world and in a large number of industrial sectors are aware of the importance of the exchange of open and well-documented high-quality data supported by reliable and effective data services [8] to transfer knowledge, boost innovation, and introduce new business models [9,10]. A number of cultural, ethical, financial, legal, and technical factors [4–6] hinder the exploitation of its full advantages. Conversely, some of these challenges can also enable public-private collaborations if enough suitable technical and organizational resources are allocated [11]. To ensure scientific progress and sustainable development, collaborative initiatives that involve both private (including academia and others as “not-for-profit organizations”, or NFPO) and private (for-profit organizations, or FPO) institutions are required [12–14]. This connection between private and public sectors forms the foundation of knowledge-based innovation systems [15], since it allows universities to know and eventually follow market needs, and support companies by providing them with additional research and development opportunities. Governmental institutions play the important role of facilitating the coordination of policies and investments. Every organization, to a greater or lesser extent, does rely on data assets and data management resources which can be used to create knowledge and evolve business processes [16]. Thus far, the public administration has lagged behind the private and research sector in terms of big data use based on the main components of business strategy, organizational capabilities, big data application strategy, and IT infrastructure, while data collaborations could further benefit from them being up to date on big-data readiness and the digital transformation [17,18]. Such cross-sector and public-private partnerships that add value to a social good or innovation appear in various forms, which have been described taxonomically by their dimensions and characteristics [19], dynamics and governance [20], and in terms of sustainability [21].

The monetary value of data is often perceived differently among stakeholders due to its complex nature, together with quality information asymmetry and lack of pricing standards [22]. The current evolution of data generation by dispersed and linked sources makes data not reliably attributable to one identifiable person. This demands collaborative strategies for data networks [23]. Generally speaking, collaborative data networks contribute to the successful generation of quality information by considering data value, management, and technology [24]. Moreover, extended data networks can increase the amount of data collected as well as interrelated models [25].

The interaction and collaboration between data producers and consumers usually occurs in so-called “data ecosystems”, i.e., environments that contain the necessary infrastructures and services for the exchange, processing, or reuse of available data [26]. Such data ecosystems can only work if the right tools for federated data management, and the access to and between data collections, are in place [27–31]. Federated data environments, a particular case of data ecosystems characterized by the control over data retained by data producers at all times (as opposed to centralizing data in a single repository from which it is retrieved for analysis), are best placed to take advantage of the benefits of data sharing data ecosystems [32], especially if they incorporate basic linked data principles, such as interlinking between data and the accompanying services, that allow the integration of data silos [33], or establish data spaces through the use of linked data frameworks [34,35]. Several examples of joint data platforms demonstrate the benefits of sharing data collections, and support the transition from private to shared domains [36–40]. A closer look at any of

these platforms shows that in order to fulfill their mission they must deal with the technical management of distributed data infrastructures together with the accompanying organizational and legal issues. Lynn et al. [41] and Grunzke et al. [42] have shown that this can be achieved by focusing on the optimization of organizational and technical requirements, management of the existing infrastructure, training, and sustainability. Other matters that require consideration are the additional information required when organizations of different types exchange data, and lastly, the human factor [43–45]. This latter aspect of data network systems is related to the development of competences and training, but is especially relevant since awareness creation, formation of a community, and the design and implementation of best practices, are of crucial importance for the success of current and future data exchange infrastructures [45,46].

The establishment of a fully functional data ecosystem across all Europe is being pursued by large initiatives like the European Open Science Cloud (EOSC), lead by the European Commission, and GAIA-X, lead by the industry. GAIA-X focuses more on the needs of companies, aiming to enable data sharing while allowing participants to retain data sovereignty [47]; EOSC is conceived mainly for research institutions. In some specific sectors, for example, health and life sciences, domain-specific data ecosystems are under development [38,48]. However, top-down initiatives should go together with complementary bottom-up approaches to overcome the numerous challenges they face. National or regional infrastructures are necessary to develop use cases that then feed into the larger initiatives, following results that show that they significantly increase innovation output [49].

3. Materials and Methods

We conducted an interview study including survey questions with regional stakeholders from various sectors and scientific domains about the current status and expected evolution of their ideal technical infrastructure.

3.1. Procedure & Analysis

The interviews were conducted online between February and July 2021 with a duration time between 30 and 70 min, using the Cisco Webex Meetings videoconferencing tool, and the resulting documents were saved on the Nextcloud server of TU Graz. Data collection followed TU Graz' data protection policies, and included informed consent information to document the interviews. The data extracted from the documents was anonymized, aggregated, and transformed into spreadsheets for further analysis. The interview template in German, a translated version of the questionnaire in English, as well as anonymized aggregated data can be found as Supplementary Material. Interview documentation related to unfeasible anonymization cannot be made public. All data is stored under restricted access, available only to IDE@S project members, and will be deleted when the project ends in 2022.

The template for interviews was created by a project core team of three researchers and was validated by several steps: First, the survey was reviewed by three project-related researchers different from the core-team, to gain a user perspective and ensure that essential topical questions have been captured. Using this feedback, the questions were refined to ensure intelligibility and avoid redundancy. Two participants, from a NFPO and FPO respectively, were chosen for a test run and feedback session in order to enhance the comprehensibility and overall evaluation possibilities of the questionnaire, by introducing additional predefined answers and scoring options for further insights. All interviews were carried out by the same researcher from the project core team to ensure the same systematic entry and documentation of answers, and a second core team member attended three sessions chosen randomly to enable further control over the process. The open questions were recorded with permission from interviewees to allow for repeated listening to answers, so that their full content could be grasped and individual viewpoints well understood.

3.2. Questionnaire

The interviews were conducted according to a questionnaire, presented to the participants while notes were taken by the interviewer. The questionnaire was structured in the following sections: Information about the interviewee's organization, introductory questions, and main questions (31 in total). The questionnaire covers different topics in data management and data science. First, aspects of technical resources and challenges are covered, followed by a survey of changes in technical infrastructure over the recent years. This leads to questions on changes in technical infrastructures planned for the next years, continued by questions on the topic of human resources for data management and data science. Some open questions are intended to get insights to the general data operation methods. Finally, there are questions on collaborative working schemes, as well as network integration.

3.3. Participants

Interviews were carried out with respondents from $n = 17$ institutions, corresponding to 9 FPOs and 8 NFPOs. Several more potential Styrian stakeholders have been asked via mail or phone to be involved in the interview study; these 17 are the number of all respondents who positively answered the request to participate in the interview study. FPOs comprise private companies from several industrial sectors, while NFPOs include institutions from the public sector, with the exception of a third-party construct of private NFPO that was formerly a public institution. NFPOs can be further classified into academia ($n = 5$ out of 8), government, or department in public administration ($n = 3$ out of 8). Stakeholder characteristics from respondents are presented in Figure 1. Participants come from several disciplines, grouped into Artificial Intelligence (AI), IT services and telecommunication, health, geosciences and geostatistics, mobility and logistics, and finance and insurance. All respondents hold leading positions or management responsibilities related to IT departments. The size of the organizations includes everything from start-ups and small working groups with up to 10 employees, small-to-medium enterprises (SMEs) between 11 and 250 employees, and large institutions with over 250 employees and up to the five-digit range.

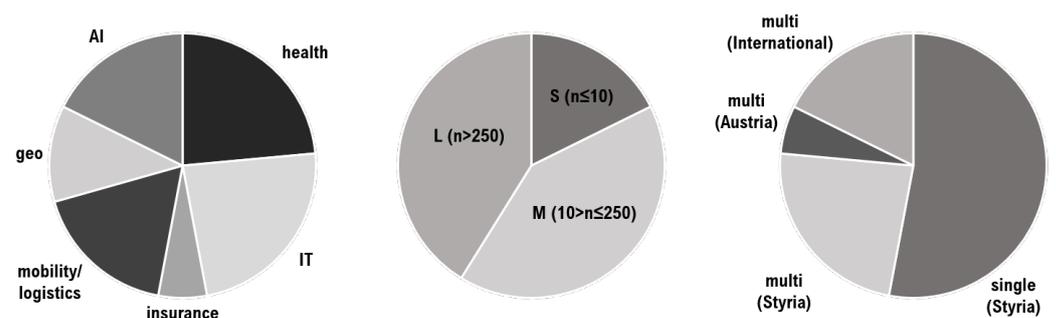


Figure 1. Participant domains (left). Size of participating organizations: Small (S), medium (M), and large (L) (center). The number of sites (single-/multi-site) are shown as a regional, national, or international organization (right). Interview-Guide Section S1 in the Supplementary Materials.

4. Results

We investigated the toward data sharing, technical infrastructure requirements, practices, and expert knowledge (skills) on collaborative data use. Results on each of these aspects are presented as quantifiable results from survey responses and summarized answers from open questions thereafter. We processed the survey responses to extract information about (i) attitudes of stakeholders towards data sharing in general, (ii) understanding data-related tasks, (iii) requirements for technical infrastructure, (iv) standard practices, (v) necessary expert knowledge on collaborative data use, and (vi) the habits of stakeholders in Styria toward data reuse and exchange in NFPOs and FPOs.

First, the different positions on data sharing of stakeholders from the three classes considered (FPOs, PA, and academia) are presented in Section 4.1. Section 4.2 shows then that the different data handling practices cannot be mainly attributed to class- or domain-specific rules, but reflect decisions made by within each organization. Data infrastructure requirements and outsourcing practices are presented in Section 4.3.

Additionally, for classification according to their public or private character, we found it useful to separate stakeholders by their domain or discipline of activity: The interviewees belonged to organizations working in AI and machine learning (ML), environmental and geosciences, health and life sciences, ICT production and IT services, finances and insurance, mobility and logistics, and smart factory. The results on domain-specific differences, shown in Section 4.4, can be found foremost in best practices and standard protocols and tools. This section also contains best practices and general standards common to all disciplines.

4.1. Different Positions on the Value of Data and Data Sharing

Figure 2 summarizes the knowledge about and attitude towards the guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets (FAIR) among interviewees. Our first finding is that, according to the responses, FAIR principles are much more familiar among public organizations (including research and administration institutions) than private companies. Even after the FAIR principles were described to respondents from private business with low familiarity with them, most of them considered that they were not relevant to their organization. This might go back to the fact that in the public sector, data sharing is a common practice, while in the private sector, data predominantly uses data internally (i.e., only within each organization). While all sectors reuse data to a certain degree, the repeated exploitation of available resources happens much more frequently in public research organizations.

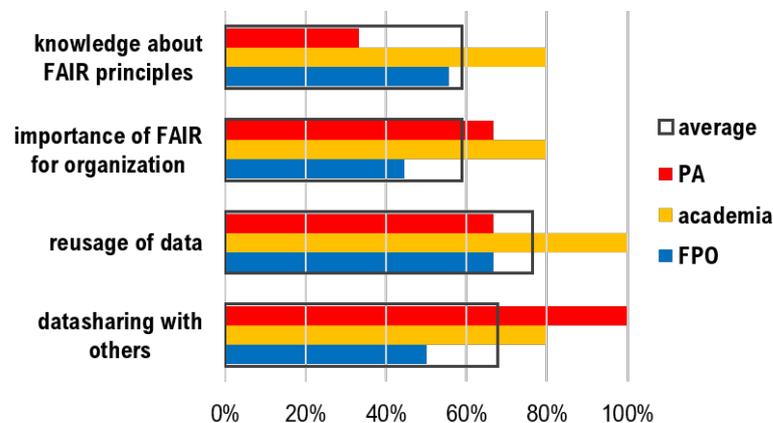


Figure 2. Attitude towards findable, accessible, interoperable, reusable (FAIR) principles among interviewees (average = gray), from for-profit organizations (FPOs) (blue), academia (yellow), and public administration (PA) institutions (red). Interview-Guide Section S4.7 in the Supplementary Materials.

4.2. Data-Related Tasks Differing between Institutions Are Not Class- or Domain-Specific

Participants were asked about their regular work routine related to data handling. Most institutions handle an average amount of data of between 1 to 100 terabytes (TB) per month, as shown in Figure 3. Data handling includes time-consuming tasks of data management, generation, retrieval, storage, analysis, and other data manipulations. The different data-related tasks in the daily-work routine are presented in Figure 4 left and show that storage is common to all sectors, whereas the usage of external cloud resources is less widespread. Next to the daily work-routine, we asked for the general utilization of common data applications. In Figure 4 right, we present the most common tools used for data management, using the options of the questionnaire. Digital documentation is a standard procedure in all organizations, while other tools such as resource scheduling,

big data techniques, and AI and ML methods, which are frequently used in academia and FPOs, are less common in the public administration sector and are employed by slightly over half of all participants.

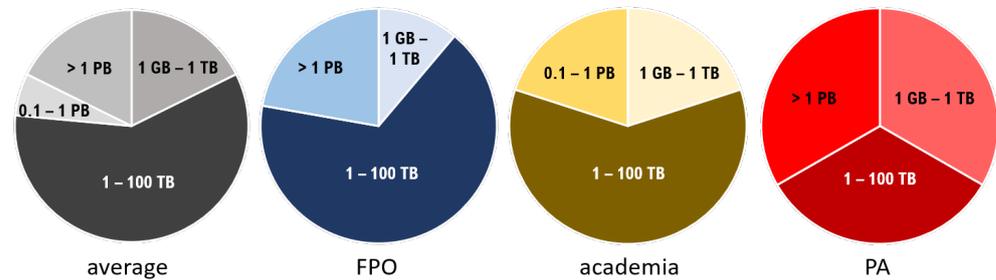


Figure 3. Average managed data per month among respondents, average (gray), FPO (blue), academia (yellow), and public administration (red), color grades relate to rankings based on frequency of responses in one category or readability. Interview-Guide Section S4.6 in the Supplementary Materials.

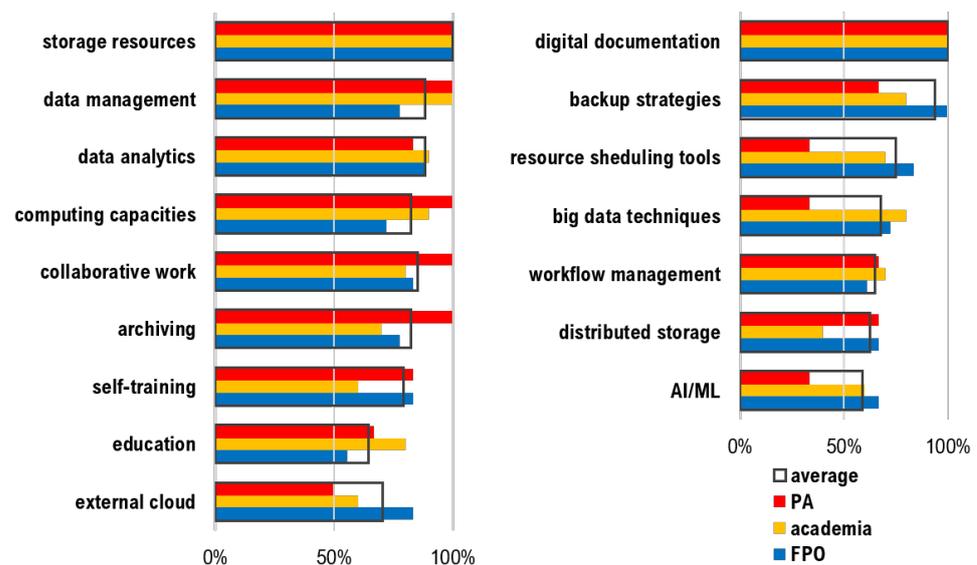


Figure 4. Data-related tasks among respondents in their daily work-routine (left). Tools used in general for data-related tasks by % organizations (right). From FPOs (blue), academia (yellow), PA institutions (red), and average (gray). Interview-Guide Sections S2.2 and S4.5 in the Supplementary Materials.

4.3. Need for Changes in Technical Infrastructure

The time evolution of data infrastructure requirements follows similar trends across all organization types, with most respondents wishing to expand and update their technical infrastructure, though the urgency of such changes is different between private and public organizations, as can be seen in Figure 5 left. Academic institutions seem to be the best equipped, and therefore express a lower pressure to upgrade their technical infrastructure, while government-related organizations in the public administration show a clear need of additional equipment and tools. The feedback on changes in data-related tasks presented in Figure 5 right shows that the need to expand data sources is present in almost 90% of organizations, and the demand for data, which has an immediate effect on the need for additional storage, appears in over 85% of them. The increasing complexity of data analysis identified by around 80% of the interviewees is a clear indicator of the need for additional computing power (70%).

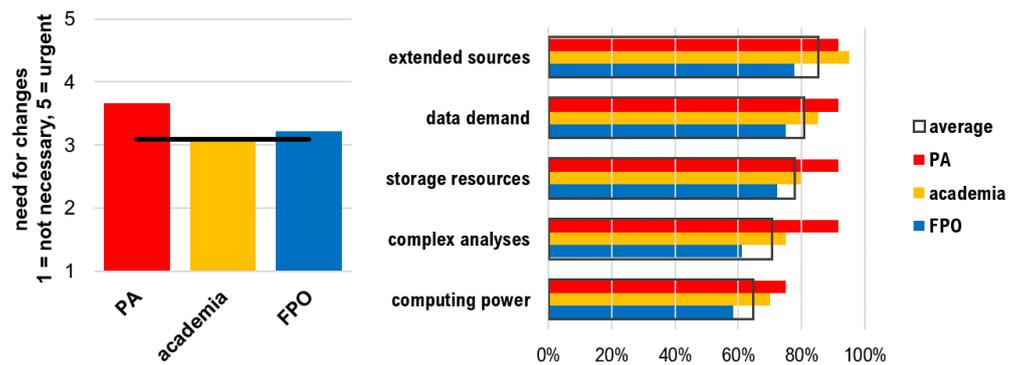


Figure 5. Need for changes in technical infrastructure (left) among interviewees (average = gray), from FPOs (blue), PA departments (red), as well academia (yellow). Change assessment of data-related tasks, applicability for % organizations (right). Interview-Guide Sections S3.2 and S3.7 in the Supplementary Materials.

The acquisition of technical solutions over the last five years is summarized in Figure 6. While most of the new equipment, tools and container solutions have been built or designed in-house, repositories and cloud services tend to be outsourced. Versioning systems, repositories and analytical tools, of primary interest to businesses and research organizations, are usually built in-house, but a considerable part of them are also outsourced. According to the responses from public administration, such types of organizations do not acquire any analytical tools or containers, but instead have designed versioning systems in-house. Only a small portion of high-performance computing (HPC) and repositories were outsourced.

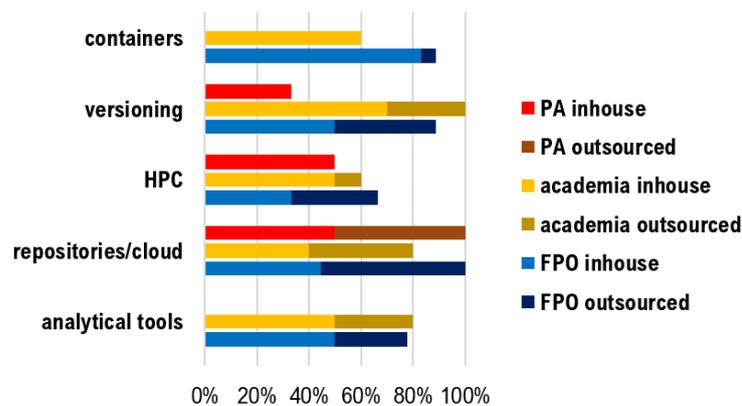


Figure 6. Recent acquisitions of technical infrastructure: equipment and tools built in-house or outsourced by % respondents, average (gray), FPOs (blue), academia (yellow), and public administration (red). Interview-Guide Section S3.3 in the Supplementary Materials.

4.4. Open Questions

Together with the closed questions already described, the questionnaire also contained open questions (i.e., allowing free text answers) about other changes needed in the data infrastructures related to their organization, like standards and interoperability, human resources, automation, storage, and computing power. This time the different fields of expertise of the interviewees were taken into account in the answers.

The questionnaire was designed to identify the daily work-routines and standard practices typical of the different domains (see Figure 4 and Section 4.2 above for domain-independent tasks). The important domain-independent (de facto) standards that came up during the conversations are listed in Table 1, which also shows best practices and common procedures. The activities indicated include organizational matters related to information technology and management systems, security techniques, and recommended tools. Stan-

dards provide a defined instruction to practices. In some of the conversations on this topic, one standard noted by many participants was the list of requirements for information security management systems provided by the International Standards Organization (ISO). The Structured Query Language (SQL) standard introduced by the American National Standards Institute (ANSI) and its variations provided by Oracle [50] and other database management systems were also mentioned. Other important standards noted frequently were the open source distributed version control system Git and SAP for electronic resource planing (ERP). Among the de facto standards, the proprietary Matlab and Simulink for model simulations in the industrial and academic engineering sector, and the well-known word and slide processing tools of the Microsoft suite were often mentioned. Next to extensible markup language (XML) as the standard for generalized markup language, Digital Imaging and Communications in Medicine (DICOM) has been indicated as the standard for information imaging.

Best practices refer to common procedures and processes established at institutions, particularly where there are not uniform domain community standards available. They include organizational practices as e.g., internal rules for documentation and life-cycle management or the introduction of a chief data officer (CDO) to units. Technical approaches comprise open source frameworks, interoperability development testing, the general application of virtual servers and firewalls, and specific applications, such as agile ML.

Table 1. Applied IT-related standards and technologies as well as best practices.

General (de facto) standards	ISO 27001, SQL & relational databases, GIT, SAP, Matlab Simulink, Microsoft Suite, XML, DICOM
Best practices	Internal rules of documentation, life-cycle management, chief data officer, open source frameworks, virtual servers, interoperability development testing, agile machine learning

Regarding human resources in data science, one in every four of all interviewees (28 %) showed a level of satisfaction with the current situation, while nearly half of them (44%) expressed a need for additional resources, while others refused to comment on this issue. Currently, as presented in Table 2, the training of employees in data science and data management, including related procedures and other administrative tasks, happens in most occasions through self-learning, complemented by introductory sessions from experienced staff. Another aspect that comes out in the responses is that people with a technical university degree are more likely to be employed to perform data-related tasks; however, the broad scope of areas touched by data science requires one to grasp concepts and know-how from several disciplines, including non-technical ones, and requires continuous training of staff.

Table 2. Demand for staff training in the data area.

Self-learning, learning by doing
Continuous training
Technical studies & discipline-specific
Certifications
Data security & data management

Regarding previous and ongoing cooperative projects, respondents included partners from various fields of public administrative bureaus, consulting services, contract research, university or business partners, and industry-coupled research projects.

Among the most relevant qualitative factors that enable cooperation in data science mentioned in the responses, as shown in Table 3, we can mention the existence of open

data policies and common standards, the organization of a sufficient number of workshops where the benefits and structure of the exchange networks can be discussed, the integration with international channels and initiatives, and the creation of enough grants for the general public. Respondents also described the ideal circumstances that would ensure future collaboration, indicated in the table as “Requirements” and “Additional factors”. Automation was mentioned with regard to infrastructures for data exchange, management, and steps of analysis, as well as routine processes. Process automation can be achieved through technologies such as the Integrated Rule-Oriented Data System (iRODS), in which rules to automate data management tasks can be implemented by so-called “rule engines” [51]. Cooperation is critical to the exchange of know-how and requires a certain degree of openness and ongoing data exchanges. Data sharing thereby comprises one-time exchange of data, and the need to reward data sharing by, for example, providing some service in return. Cooperation helps to create visibility of available data and possible project partners, but it also presupposes it.

Table 3. Fostering data exchange and cooperation.

Cooperation facilitators	open data policies, common standard, workshops, integration with international channels, general public grants
Requirements	automation, computing power, human resource, standards & interoperability, structure instead of quantity
Additional factors for collaborative data use	openness & collaborative thinking, permanent exchange, exchange of know-how, service in return for data provision, visibility of own data

5. Discussion

The goal of this study is to highlight the factors that facilitate the cooperative use of data and help understand the attitudes of Styrian stakeholders towards data exchange and cooperation. The main elements investigated include the technical requirements as well as specialized human resources and IT practices that are necessary for given data-related tasks. The wide variability observed in the responses from public institutions to questions on the use of data, sharing practices, and technical requirements, not seen in the private sector, suggested to subdivide the public group further into academia and PA. Three main messages can be extracted from the responses: Firstly, it strikes as evident that private and public institutions alike require the development of e-infrastructures for data capturing, processing, analysis, and storage to enable that enough data is shared; secondly, data management represents an essential task of the daily work-routine in all domains, but this does not necessarily imply the employment of big data and ML methods; and thirdly, the storage capacity needs frequent updating due to the fast growth of data volumes, and this requires well-thought data structuring and alternative approaches to develop cooperative thinking on data solutions. So far, the need for extended data infrastructures is managed differently among the various stakeholders, especially regarding service sourcing. Competitiveness notably present in private organizations makes this sector tend to a regular market analysis for strategic outsourcing [52]. This can be observed e.g., in the use of container-based technology, which has recently become popular for flexible application management and deployment [53]. Containers were mentioned particularly by respondents from private organizations and, in the public sector, by academia, but not by administration.

The readiness to reuse data implies that some sort of data exchange with another stakeholder is likely to take place. Our survey shows that the disposition to share data is very different across the range of disciplines and organizations to whom the interviewees belong, as well as between the private and public sector. Private companies point to two hindering factors to data exchange: First, companies harvest data for their exclusive use and thus keep a competitive advantage over current or potential business rivals, according to their ultimate goal of profit maximization. Second, the willingness to share data is lower

when the benefits it can bring to the general public (i.e., the society) are not clear, or if they clash with the interests of a company. In the public sector, open data policies have already been applied, but they still need both top-down and bottom-up actions to be successfully implemented on a large scale. Such changes can take place in the private sector as well, especially if introduced first in small- and medium-sized enterprises, where they should be easier to implement than in large businesses. The big difference in the degree of and the attitude to data sharing in several disciplines has been shown before, and an enforced surveillance of data sharing has been suggested as counteraction [54]. Data exchange can lower the barriers towards data sharing still placed by private and public organizations. It is one of the pillars on which the ongoing digital transformation relies, leading in many cases to new revenue sources and growth in an increasingly digital world. Digitization, especially when accompanied by (re)use of data, enables to automate calculations and analysis for planning and discovery, suggests future directions for research, and ultimately leads to an increase in the efficiency and quality of development.

The high frequency in which data-related tasks appear in survey responses shows that collaborative work is very important for private and public organizations, but especially for public administration, while big data techniques are predominantly used by academia and FPOs. The various applications and benefits of big data, including current open research directions based on management and manipulation of large data sets, draw the attention of organizations and researchers because of the potential delivery of insights that may be turned into added value. The possible uses of data do not end with data sharing, of course. For example, there have been increased efforts to share ML models in federated learning (FL) settings, consistent with the interest in use (and reuse) of data not only to extract knowledge, but also to be able to learn (i.e., make predictions) from them. In FL, ML algorithms are trained collaboratively across many clients (e.g., mobile phones), while the data remain decentralized, which is essential to preserve privacy in applications of ML technology [55]. With the introduction of ML, data exchange moves to a more abstract level, since data here is not explicitly exchanged, but rather only used to enable collective learning. This satisfies the ultimate aim of working with data: To obtain added value from data from multiple owners. With data sets spread over several owners and locations, one can also profit from “transfer learning”, where a model is trained on one data set (i.e., from a specific data owner), to be then applied to data owned by others. The technique can be used in both supervised and unsupervised scenarios [56]. This idea has been addressed by the European Commission in several funding calls and research initiatives, including studies to learn from multiple sets of proprietary data across numerous European stakeholders [57].

In line with findings by the European Commission, who has identified interoperability frameworks as critical components for joint undertakings [58], several responses from interviewees have highlighted interoperability as a requirement to foster data exchange and cooperation. In practice, this involves the introduction of standards and interoperability for both technical and data standards. Data standards that follow domain-specific formats and practices must be developed in collaboration with experts of the respective domains, whereas technical standards, which are related to the underlying technical requirements, need to be domain-agnostic. There have been recent approaches to develop metadata standards for multi-disciplinary data [59], which would undoubtedly enhance data interoperability, but this remains a difficult task in most domains. A possible approach to the standardization of data exchange is shown by recent advances in medical informatics, where it has been possible to develop a standard that enables interoperability between healthcare systems [60]. The format (named “Health Level 7 standard Fast Healthcare Interoperability Resources”, or “HL7 FHIR”) is modular and contains information on identity, metadata, human readable summaries, and standard data. In a European context, where very different languages are used simultaneously, such standards can provide both language-independent as well as language-specific layers to support data exchange.

Another essential factor that can determine whether collaborative data infrastructures are successful is the transfer of know-how among staff. As noted by respondents from all

sectors, training in data science and for related tasks happens currently via self-learning in most cases. Due to the fast evolution in data science and e-infrastructures, continuous training and education will be necessary to create and keep the competences of staff at all types of institutions in this era. Here it is worth mentioning the digitization support to companies provided by the Digital Innovations Hubs (DIH) programme of the European Commission. There are currently around 400 fully operational DIHs in computing, digital twins, AI, and big data in the EU, as listed in the Joint Research Centre platform, with more expected to come in the second quarter of 2022 [61,62]. The role of DIHs is to organize networking events and workshops, and provide technological support for the exchange of know-how. They have become fundamental in connecting stakeholders with each other, and increasing the information flow on regional infrastructures. This is consistent with the idea that fostering collaboration depends on forging links, creating new contacts, and facilitating frequent information exchange.

In the long run, it is clear that one initiative alone will not be able to help overcome all challenges faced by research institutions and businesses in the digital age, and synergies between projects, platforms, and technological as well as organizational solutions are needed. Collaboration between individual projects will broaden and strengthen participation in large-scale initiatives.

6. Conclusions

Collaborative efforts that involve public and private entities facilitate innovation and are crucial to scientific progress and technological development. Data have emerged as the most important assets to realize collaborations. We therefore investigated factors influencing the collaborative use of data among various stakeholders from private, as for-profit organizations, and public institutions, that we further divided into academia and public administration. Results indicate that while there is a will from all stakeholders to engage in collaborative data spaces, the organizational changes and improvements required to reach the necessary standards that would make this possible still lie in the future. We have found that while both public and private entities could benefit from, and are interested in accessing data sources other than their own, existing and undiscovered synergies between isolated or fragmented systems still have to be exploited in order to achieve data interoperability and enable cooperation among institutions. The isolated factors influencing data collaborations have to be addressed individually and help to understand data-driven value creation. If the demands from different stakeholders can be properly addressed, it is possible to envision a future where data exchange across disciplines becomes a reality, with the potential of leading to true knowledge discovery. Our survey, performed within the project IDE@S in the Austrian federal state of Styria, presents a regional perspective that could be further translated to a wider context as applicable to other places. The regional aspect of this work represents a strength but also one limitation in regard to translating noted key points to a European level. Additionally, in order to highlight domain-specific issues, such as the different types of data used, future research is necessary on select individual use cases, or involving a much higher number of participants in similar studies. A large-scale survey could also give an overview of regional or national digitization progress, and furthermore, would allow for a statistical analysis of elementary factors influencing data collaboratives. Furthermore, future studies in the course of the IDE@S project will identify training gaps and highlight additional or supplementary courses and curricula that may need to be introduced for prospective joint data infrastructures. Additionally, investigating use cases will also allow one to understand the peculiarities of domain-specific data collaboratives.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/data7020020/s1>, Interview questionnaire in German, translated questionnaire in English, aggregated data from anonymized samples.

Author Contributions: Conceptualization, S.S., M.R.M., and C.J.-Q.; methodology, S.S., M.R.M., and C.J.-Q.; formal analysis, C.J.-Q.; data curation, C.J.-Q.; writing—original draft preparation, C.J.-Q., S.S., and M.L.; writing—review and editing, M.R.M.; visualization, C.J.-Q.; supervision, S.S.; project administration, M.R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Amt der Steiermärkischen Landesregierung, grant number ABT08-24920/2020-14.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all interview candidates involved in the study.

Data Availability Statement: The data presented in this study are available in aggregated form as supplementary material.

Acknowledgments: We thank all participants for the provided information, discussions, and their interest in the project. Open Access Funding by the Graz University of Technology.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
ANSI	America National Standards Institute
CDO	Chief data officer
DIH	Digital Innovation Hub
EOSC	European Open Science Cloud
ERP	Electronic resource planning
FAIR	Findable Accessible Interoperable Reusable
FL	Federated Learning
FPO	For-profit organization
GB	Gigabyte
HR	Human resources
IDE@S	Innovative Data Environment @ Styria
iRODS	integrated Rule-Oriented Data System
ISO	International organization for Standardization
ML	Machine learning
NFPO	Not-for-profit organization
PA	Public administration
PB	Petabyte
SQL	Structured query language
TB	Terrabyte
XML	extensible markup language

References

1. Chatterjee, S.; Chaudhuri, R.; Vrontis, D. Does data-driven culture impact innovation and performance of a firm? An empirical examination. *Ann. Oper. Res.* **2021**, *1*–26. [[CrossRef](#)]
2. Monino, J.L. Data value, big data analytics, and decision-making. *J. Knowl. Econ.* **2021**, *12*, 256–267. [[CrossRef](#)]
3. Aubin, I.; Cardou, F.; Boisvert-Marsh, L.; Garnier, E.; Strukelj, M.; Munson, A.D. Managing data locally to answer questions globally: The role of collaborative science in ecology. *J. Veg. Sci.* **2020**, *31*, 509–517. [[CrossRef](#)]
4. Mannheimer, S.; Pienta, A.; Kirilova, D.; Elman, C.; Wutich, A. Qualitative data sharing: Data repositories and academic libraries as key partners in addressing challenges. *Am. Behav. Sci.* **2019**, *63*, 643–664. [[CrossRef](#)]
5. Fernandes, G.; O’Sullivan, D.; Ferreira, L.M.D. Addressing the Challenges to Successfully Manage University- Industry R&D Collaborations. *Procedia Comput. Sci.* **2022**, *196*, 724–731.
6. Zhang, Z.; Song, S.; Yu, J.; Zhao, W.; Xiao, J.; Bao, Y. The elements of data sharing. *Genom. Proteom. Bioinform.* **2020**, *18*, 1. [[CrossRef](#)]
7. Fafalios, P.; Petrakis, K.; Samaritakis, G.; Doerr, K.; Kritsotaki, A.; Tzitzikas, Y.; Doerr, M. FAST CAT: Collaborative Data Entry and Curation for Semantic Interoperability in Digital Humanities. *arXiv* **2021**, arXiv:2105.13733.

8. Cios, K.J.; Swiniarski, R.W.; Pedrycz, W.; Kurgan, L.A. The knowledge discovery process. In *Data Mining*; Springer: Boston, MA, 2007; pp. 9–24.
9. Attard, J.; Orlandi, F.; Auer, S. Data value networks: Enabling a new data ecosystem. In Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, USA, 13–16 October 2016; pp. 453–456.
10. Kitsios, F.; Papachristos, N.; Kamariotou, M. Business models for open data ecosystem: Challenges and motivations for entrepreneurship and innovation. In Proceedings of the 2017 IEEE 19th Conference on Business Informatics (CBI), Thessaloniki, Greece, 24–27 July 2017; Volume 1, pp. 398–407.
11. Sjö, K.; Hellström, T. University–industry collaboration: A literature review and synthesis. *Ind. High. Educ.* **2019**, *33*, 275–285. [[CrossRef](#)]
12. Carbonara, N.; Pellegrino, R. The role of public private partnerships in fostering innovation. *Constr. Manag. Econ.* **2020**, *38*, 140–156. [[CrossRef](#)]
13. Davis, A.M.; Engkvist, O.; Fairclough, R.J.; Feierberg, I.; Freeman, A.; Iyer, P. Public-Private Partnerships: Compound and Data Sharing in Drug Discovery and Development. *SLAS Discov. Adv. Sci. Drug Discov.* **2021**, *26*, 604–619. [[CrossRef](#)]
14. Sergi, B.S.; Popkova, E.G.; Borzenko, K.V.; Przhedetskaya, N.V. Public–private partnerships as a mechanism of financing sustainable development. In *Financing Sustainable Development*; Palgrave Studies in Impact Finance; Springer: Cham, Switzerland, 2019. [[CrossRef](#)]
15. Leydesdorff, L. The triple helix model and the study of knowledge based innovation systems. *arXiv* **2009**, arXiv:0911.4291.
16. Hannila, H.; Silvola, R.; Harkonen, J.; Haapasalo, H. Data-driven begins with DATA; potential of data assets. *J. Comput. Inf. Syst.* **2022**, *62*, 29–38. [[CrossRef](#)]
17. Klievink, B.; Romijn, B.J.; Cunningham, S.; de Bruijn, H. Big data in the public sector: Uncertainties and readiness. *Inf. Syst. Front.* **2017**, *19*, 267–283. [[CrossRef](#)]
18. Craglia, M.; Micheli, M.; Hradec, J.; Calzada, I.; Luitjens, S.; Ponti, M.; Scholten, H.J.; Boter, J. *Digitranscope: The Governance of Digitally-Transformed Society*; EUR 30590; Publications Office of the European Union: Luxembourg, 2021. [[CrossRef](#)]
19. Susha, I.; Janssen, M.; Verhulst, S. Data collaboratives as a new frontier of cross-sector partnerships in the age of open data: Taxonomy development. In Proceedings of the 50th Hawaii International Conference on System Sciences, Hilton Waikoloa Village, HI, USA, 4–7 January 2017.
20. Susha, I.; Gil-Garcia, J.R. A collaborative governance approach to partnerships addressing public problems with private data. In Proceedings of the 52nd Hawaii International Conference on System Sciences, Maui, HI, USA, 8–11 January 2019.
21. Martin, S.; Gautier, P.; Turki, S.; Kotsev, A. *Establishment of Sustainable Data Ecosystems*; EUR 30626 EN; Publications Office of the European Union: Luxembourg, 2021. [[CrossRef](#)]
22. Tang, Q.; Shao, Z.; Huang, L.; Yin, W.; Dou, Y. Identifying Influencing Factors for Data Transactions: A Case Study from Shanghai Data Exchange. *J. Syst. Sci. Syst. Eng.* **2020**, *29*, 697–708. [[CrossRef](#)]
23. Kilic, D.; Crabtree, A.; McGarry, G.; Goulden, M. The cardboard box study: Understanding collaborative data management in the connected home. *Pers. Ubiquitous Comput.* **2022**, *26*, 155–176. [[CrossRef](#)]
24. Chen, Y.C.; Lee, J. Collaborative data networks for public service: Governance, management, and performance. *Public Manag. Rev.* **2018**, *20*, 672–690. [[CrossRef](#)]
25. Lazarova-Molnar, S.; Mohamed, N. Collaborative data analytics for smart buildings: Opportunities and models. *Clust. Comput.* **2019**, *22*, 1065–1077. [[CrossRef](#)]
26. Oliveira, M.I.S.; Lima, G.d.F.B.; Lóscio, B.F. Investigations into Data Ecosystems: A systematic mapping study. *Knowl. Inf. Syst.* **2019**, *61*, 589–630. [[CrossRef](#)]
27. Schmitt, C.; Buchinal, M.R. Data management practices for collaborative research. *Front. Psychiatry* **2011**, *2*, 47. [[CrossRef](#)]
28. Mandl, K.D.; Glauser, T.; Krantz, I.D.; Avillach, P.; Bartels, A.; Beggs, A.H.; Biswas, S.; Bourgeois, F.T.; Corsmo, J.; Dauber, A.; et al. The Genomics Research and Innovation Network: Creating an interoperable, federated, genomics learning system. *Genet. Med.* **2020**, *22*, 371–380. [[CrossRef](#)]
29. Hohlenweger, T.S.; Pinheiro, M.; Spínola, A.L.; Macedo, I.L.; Martins, J.S.; Dourado, R.A.; Monteiro, J.A.S. An iRODS-based distributed and federated data repository for a Multi-CMF Network for Experimentation. In Proceedings of the 2015 Latin American Network Operations and Management Symposium (LANOMS), João Pessoa, Brazil, 1–3 October 2015; pp. 62–68.
30. Fernández-del Castillo, E.; Scardaci, D.; García, Á.L. The EGI federated cloud e-infrastructure. *Procedia Comput. Sci.* **2015**, *68*, 196–205. [[CrossRef](#)]
31. Blümm, M.; Schmunk, S. Digital Research Infrastructures: DARIAH. In *3D Research Challenges in Cultural Heritage II*; Springer: Cham, Switzerland, 2016; pp. 62–73.
32. Otto, B.; Lis, D.; Jürjens, J.; Cirullies, J.; Howar, F.; Meister, S.; Spiekermann, M.; Pettenpohl, H.; Möller, F.; Rehof, J.; et al. *Data Ecosystems. Conceptual Foundations, Constituents and Recommendations for Action*; Fraunhofer ISST: Dortmund, Germany, 2019; ISSN 0943–1624.
33. Speiser, S.; Harth, A. Taking the lids off data silos. In Proceedings of the 6th International Conference on Semantic Systems, Graz, Austria, 1–3 September 2010; pp. 1–4.
34. Bizer, C. The emerging web of linked data. *IEEE Intell. Syst.* **2009**, *24*, 87–92. [[CrossRef](#)]

35. Bader, S.; Pullmann, J.; Mader, C.; Tramp, S.; Quix, C.; Müller, A.W.; Akyürek, H.; Böckmann, M.; Imbusch, B.T.; Lipp, J.; et al. The International Data Spaces Information Model—An Ontology for Sovereign Exchange of Digital Content. In *International Semantic Web Conference*; Springer: Cham, Switzerland, 2020; pp. 176–192.
36. Talbot, S.R.; Bruch, S.; Kießling, F.; Marscholke, M.; Jandric, B.; Tolba, R.H.; Bleich, A. Design of a joint research data platform: A use case for severity assessment. *Lab. Anim.* **2020**, *54*, 33–39. [[CrossRef](#)]
37. Mohr, C.; Friedrich, A.; Wojnar, D.; Kenar, E.; Polatkan, A.C.; Codrea, M.C.; Czermel, S.; Kohlbacher, O.; Nahnsen, S. qPortal: A platform for data-driven biomedical research. *PLoS ONE* **2018**, *13*, e0191603. [[CrossRef](#)] [[PubMed](#)]
38. Harrow, J.; Drysdale, R.; Smith, A.; Repo, S.; Lanfear, J.; Blomberg, N. ELIXIR: Providing a sustainable infrastructure for life science data at European scale. *Bioinformatics* **2021**, *37*, 2506–2511. [[CrossRef](#)] [[PubMed](#)]
39. Bahmani, A.; Alavi, A.; Buerger, T.; Upadhyayula, S.; Wang, Q.; Ananthakrishnan, S.K.; Alavi, A.; Celis, D.; Gillespie, D.; Young, G.; et al. A scalable, secure, and interoperable platform for deep data-driven health management. *Nat. Commun.* **2021**, *12*, 1–11. [[CrossRef](#)]
40. Wieser, F.; Stryeck, S.; Lang, K.; Hahn, C.; Thallinger, G.; Feichtinger, J.; Hack, P.; Stepponat, M.; Merchant, N.; Lindstaedt, S.; et al. A local platform for user-friendly FAIR data management and reproducible analytics. *J. Biotechnol.* **2021**, *341*, 43–50. [[CrossRef](#)]
41. Lynn, T.; Mooney, J.G.; Domaschka, J.; Ellis, K.A. *Managing Distributed Cloud Applications and Infrastructure: A Self-Optimising Approach*; Springer: Heidelberg, Germany, 2020.
42. Grunzke, R.; Adolph, T.; Biardzki, C.; Bode, A.; Borst, T.; Bungartz, H.J.; Busch, A.; Frank, A.; Grimm, C.; Hasselbring, W.; et al. Challenges in creating a sustainable generic research data infrastructure. *Softwaretechnik-Trends* **2017**, *37*, 74–77.
43. Mikalef, P.; Krogstie, J. Investigating the data science skill gap: An empirical analysis. In Proceedings of the 2019 IEEE Global Engineering Education Conference (EDUCON), Dubai, United Arab Emirates, 8–11 April 2019; pp. 1275–1284.
44. Miller, S.; Hughes, D. *The Quant Crunch: How the Demand for Data Science Skills Is Disrupting the Job Market*; Burning Glass Technologies: Boston, MA, USA, 2017.
45. Poole, A.H. How has your science data grown? Digital curation and the human factor: A critical literature review. *Arch. Sci.* **2015**, *15*, 101–139. [[CrossRef](#)]
46. Jerman, A.; Pejić Bach, M.; Aleksić, A. Transformation towards smart factory system: Examining new job profiles and competencies. *Syst. Res. Behav. Sci.* **2020**, *37*, 388–402. [[CrossRef](#)]
47. Braud, A.; Fromentoux, G.; Radier, B.; Le Grand, O. The Road to European Digital Sovereignty with Gaia-X and IDSA. *IEEE Netw.* **2021**, *35*, 4–5. [[CrossRef](#)]
48. Iacob, N.; Simonelli, F. Towards a European Health Data Ecosystem. *Eur. J. Risk Regul.* **2020**, *11*, 884–893. [[CrossRef](#)]
49. Lopes, J.; Franco, M. Review about regional development networks: An ecosystem model proposal. *J. Knowl. Econ.* **2019**, *10*, 275–297. [[CrossRef](#)]
50. Petković, D. JSON integration in relational database systems. *Int. J. Comput. Appl.* **2017**, *168*, 14–19. [[CrossRef](#)]
51. Fortner, B.; Ahalt, S.; Coposky, J.; Fecho, K.; Krishnamurthy, A.; Moore, R.; Rajasekar, A.; Schmitt, C.; Schroeder, W. *Control Your Data: iRODS, the Integrated Rule-Oriented Data System*; University of North Carolina at Chapel Hill: Chapel Hill, NC, USA, 2014.
52. Taponen, S.; Kauppi, K. Service outsourcing decisions—A process framework. *J. Glob. Oper. Strateg. Sourc.* **2020**, *13*, 171–194. [[CrossRef](#)]
53. Rodriguez, M.A.; Buyya, R. Container-based cluster orchestration systems: A taxonomy and future directions. *Softw. Pract. Exp.* **2019**, *49*, 698–719. [[CrossRef](#)]
54. Tedersoo, L.; Küngas, R.; Oras, E.; Köster, K.; Eenmaa, H.; Leijen, Ä.; Pedaste, M.; Raju, M.; Astapova, A.; Lukner, H.; et al. Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **2021**, *8*, 1–11. [[CrossRef](#)]
55. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.A.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and Open Problems in Federated Learning. *arXiv* **2019**, arXiv:1912.04977.
56. Lovrić, M.; Đuričić, T.; Tran, H.T.N.; Hussain, H.; Lacić, E.; Rasmussen, M.A.; Kern, R. Should We Embed in Chemistry? A Comparison of Unsupervised Transfer Learning with PCA, UMAP, and VAE on Molecular Fingerprints. *Pharmaceuticals* **2021**, *14*, 758. [[CrossRef](#)]
57. David, L.; Arús-Pous, J.; Karlsson, J.; Engkvist, O.; Bjerrum, E.J.; Kogej, T.; Kriegl, J.M.; Beck, B.; Chen, H. Applications of Deep-Learning in Exploiting Large-Scale and Heterogeneous Compound Data in Industrial Pharmaceutical Research. *Front. Pharmacol.* **2019**, *10*, 1303. [[CrossRef](#)]
58. Borgogno, O.; Colangelo, G. Data sharing and interoperability: Fostering innovation and competition through APIs. *Comput. Law Secur. Rev.* **2019**, *35*, 105314. [[CrossRef](#)]
59. Rumble, J.; Broome, J.; Hodson, S. Building an international consensus on multi-disciplinary metadata standards: A codata case history in nanotechnology. *Data Sci. J.* **2019**, *18*, 12. [[CrossRef](#)]
60. Bender, D.; Sartipi, K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal, 20–22 June 2013; pp. 326–331. [[CrossRef](#)]
61. Joint Research Centre. *The S3 Platform of Smart Specialisation*; European Commission: Seville, Spain, 2011.
62. McCann, P.; Ortega-Argilés, R. Smart specialisation in European regions: Issues of strategy, institutions and implementation. *Eur. J. Innov. Manag.* **2014**, *17*, 409–427. [[CrossRef](#)]