

Article

Techniques to Improve B2B Data Governance Using FAIR Principles

Cristina Georgiana Calancea *  and Lenuța Alboaie *

Faculty of Computer Science, Alexandru Ioan Cuza University, 6600 Iași, Romania

* Correspondence: gcalancea@info.uaic.ro (C.G.C.); adria@info.uaic.ro (L.A.); Tel.: +40-720-711-265 (C.G.C.)

Abstract: Sharing data along the economic supply/demand chain represents a catalyst to improve the performance of a digitized business sector. In this context, designing automatic mechanisms for structured data exchange, that should also ensure the proper development of B2B processes in a regulated environment, becomes a necessity. Even though the data format used for sharing can be modeled using the open methodology, we propose the use of FAIR principles to additionally offer business entities a way to define commonly agreed upon supply, access and ownership procedures. As an approach to manage the FAIR modelled metadata, we propose a series of methodologies to follow. They were integrated in a data marketplace platform, which we developed to ensure they are properly applied. For its design, we modelled a decentralized architecture based on our own blockchain mechanisms. In our proposal, each business entity can host and structure its metadata in catalog, dataset and distribution assets. In order to offer businesses full control over the data supplied through our system, we designed and implemented a sharing mechanism based on access policies defined by the business entity directly in our data marketplace platform. In the proposed approach, metadata-based assets sharing can be done between two or multiple businesses, which will be able to manually access the data in the management interface and programmatically through an authorized data point. Business specific transactions proposed to modify the semantic model are validated using our own blockchain based technologies. As a result, security and integrity of the FAIR data in the collaboration process is ensured. From an architectural point of view, the lack of a central authority to manage the vehiculated data ensures businesses have full control of the terms and conditions under which their data is used.

Keywords: B2B data governance mechanisms; methodologies to model and manage FAIR data; decentralized B2B data marketplace architecture; metadata as blockchain assets; transactions controlled semantic model; access policy based data sharing



Citation: Calancea, C.G.; Alboaie, L. Techniques to Improve B2B Data Governance Using FAIR Principles. *Mathematics* **2021**, *9*, 1059. <https://doi.org/10.3390/math9091059>

Academic Editors:
Octavian Dospinescu and Juan
Jose García-Machado

Received: 30 March 2021

Accepted: 7 May 2021

Published: 9 May 2021

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



Copyright: © 2021 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. Cross-Sector B2B Data Sharing and Its Impact on the Companies Ecosystem

For a long time, companies strictly relied on traditional Business-to-Business (B2B) transactions to evolve. These transactions refer to purchasing and selling physical raw goods, with the goal to complete the product manufacture process [1]. Usually, the obtained product represents the base of Business-to-Consumer (B2C) transactions. In order to ensure the success of B2C transactions, all companies involved in the B2B collaboration chain need to have an overview of the raw good demand and their supply capacity in comparison to similar businesses [2]. As a result, data sharing between companies has become a key aspect in the process of growing business opportunities in the past years. B2B data sharing refers to “making data available to or accessing data from other companies for business purposes” [3] either for free or by making a payment to the data holder. The business owner has the option to choose who to share the data with and under which conditions.

In [4], the author emphasizes the need to share data between well-established companies in order to encourage innovation. Data sharing comes as a prerequisite since one company alone cannot envision the complete perspective of the economic supply/demand

chain. When data from multiple businesses is gathered through this process, customer experience in relation to several products can also be easily identified. This helps business owners adjust their plans and products to fulfill customer needs and expectancies. Companies which imply themselves in B2B processes with their suppliers and their buyers will be able to see issues in new and existing business initiatives from multiple perspectives. As a result, business representatives will understand each other's difficulties in proposed growth plans and will focus on finding solutions that benefit everyone.

In order to achieve a performant business model, designing efficient processes of B2B data sharing is crucial. One of the main encountered difficulties in their implementation is given by the dependency between the business users and the technological department. The latter is required for providing data adapters with various custom logic, which depends on the data provider. Another issue in B2B data sharing implementation appears because of the data formats diversity, protocols and standards through which each provider chooses to deliver its data. On a platform specifically designed for the B2B data sharing use case, a large number of transactions over the datasets are performed concurrently by the collaborating businesses. Subsequently, the need for implementing synchronization and consensus mechanisms in the data sharing process appears [5,6].

Companies which engage in B2B data sharing processes tend to estimate the effort to integrate custom mechanisms in the existing business model, in order to assess the feasibility of the whole process. In the B2B perspectives report [3], redacted by the European Commission, some categories of techniques and strategies for the exchange of shared data can be outlined:

- Data monetization—companies are willing to share part of their data in order to increase their business revenues; according to a Gartner report on “Magic Quadrant for Analytics and Business Intelligence Platforms” [7], several corporations consider data sharing for profit an important part of their business strategy, amongst which Microsoft and Tableau are leaders, Oracle and Salesforce are visionaries, while IBM and Alibaba Cloud can be considered niche players; there are also business such as MicroStrategy and Looker who constantly challenge the consecrated techniques for B2B, which leads to improvements and new discoveries;
- Public data marketplaces—rely on public trusted entities that connect both data sellers and data buyers in one environment; usually, a transaction fee is perceived for all exchanges in order to keep the platform alive;
- Industrial data platforms—a secure and private environment which is restricted to group of companies exchanging data for free voluntarily in order to facilitate new product and services development;
- Technical enablers—businesses which specialize in creating data sharing flows custom for companies;
- Open data policy—companies sharing part of their data in a completely open manner;

From a technical point of view, companies involved in creating public data marketplaces, industrial data platforms and technical enablers offer solutions which intermediate the B2B data sharing process. We analyzed the most representative technical solutions with their advantages and disadvantages.

A major player which monetizes B2B data sharing processes through its public data marketplace is DataPace [8]. This company offers a decentralized global-scale marketplace to facilitate the trade of data collected from a variety of sensors and IoT devices. Businesses can sell and buy sensor's collected data through DataPace in a secure environment powered by the hyperledger fabric blockchain technology, similar to e-store web sites. Some advantages to this approach are the ability to tokenize the value of data (through platform specific tokens), to ensure its integrity through stored hashes in the blockchain and to enable smart contracts capabilities. The stakeholders of this platform are the data sellers, data buyers and validators. The validators are entities responsible to validate blocks created by the other stakeholders in the network, through their transactions. Network security is ensured through the PBFT consensus algorithm, which ensures the system

has fault tolerance to 1/3 nodes becoming malicious. The actual data sharing process is facilitated by smart contracts defined in the system between companies, which contain the set of conditions under which data is exchanged. This approach ensures other transaction costs associated with creating a traditional law contract are reduced and the security of the transaction is enhanced. This solution is highly scalable, characteristic which is assured by the underlying decentralized architecture. A strong downside to this solution is that there are no privacy enhancing mechanisms included by default. This aspect must be ensured by companies developing the smart contracts. Data governance mechanisms are not available by default, which may result in issues of ownership and control of usage over the shared data. Finally, the shared data is not annotated to be reusable and interoperable, which may lead to redundancy issues and the need to build custom integrations for each integration within a B2B process. Another limitation of this platform is that the shared data is strictly collected from IoT sensors and systems, which limits its usage in other B2B contexts.

A trending and emerging technical enabler for B2B data sharing is proposed by Epimorphics [9], which provides a suite of services and tools for linked data management. These services offer support for commercial customers to model their data and build applications relying on the obtained linked datasets. They also offer consultancy services with customizations depending on each business's requests. The data management platform has a three-tiered architecture, comprised of load balancing and routing, application services and storage. Load balancing is ensured by providing a company with multiple RDF servers, which provide access through API and user interface to the linked data belonging to the business. Linked data accessible through API is updated and maintained through a data management service, which is responsible with the conversion of unstructured to structured data. Since the conversion process is specific to each business, Epimorphics develops different convertors for each business in order to maximize the findability, accessibility, interoperability and reusability of their data, according to the FAIR principles. Epimorphics's business model focuses on providing support for companies to prepare their data for integration in B2B processes, rather than facilitate discovery of linked data and enabling new partnerships between companies. A significant downside to this solution is the lack of privacy centered data curation, which may lead to businesses subsequently sharing private user data. This leaves the business exposed to legal issues in terms of compliance with data confidentiality laws, such as GDPR. Another downside to this approach is the lack of integrated data governance mechanisms. Companies are able to share their datasets with their partners by providing access directly to an RDF API, authenticated through OAuth2 tokens. Nonetheless, API security through OAuth2 authentication does not guarantee that data usage can be controlled or its ownership can be asserted by the owner company.

A promising technology offering an integrated public data marketplace is iGrant.io [10]. It is a cloud-based data exchange and consent mediation platform, which targets to help businesses monetize the personal data they collect about users, while preserving user privacy rights granted through GDPR and other data protection regulations. In comparison with the other B2B data platforms, iGrant.io offers data collection mechanisms for products which are user-centered. Business customers can closely manage their data within a wallet, controlling how their data is used and shared by the company they provided access to. In turn, companies can define their own terms of data usage in the available enterprise management platform, which are then submitted to users for review and consent. These features are facilitated by a commercial software named MyData Operator, which companies can integrate in their software services. The operator provides business processes with access to an own indexed metadata registry, which stores references to the data in a distributed ledger. Using this registry, businesses can improve their own products and services by integrating datasets of user-consented data. A downside of the data exchange process facilitated by iGrant.io is that data governance is only addressed from the consumer's point of view, which can choose third party companies to share its data with. No specific mechanisms are developed to ensure companies have control over the data they make available in the registry, nor can they claim ownership and control

access rights over it. Other aspects to be improved are the interoperability and reusability of the data, which are not included in the current version of the software. The possibility to create formal agreements in the B2B sharing process between companies is also needed, in order to further develop privacy compliant enterprise data governance mechanisms.

Our proposal, which we called DataShareFair, aims to provide a viable solution, in respect to data governance principles, for the commonly encountered issues of the B2B data sharing strategies. It provides the advantages of the public data marketplaces, while also including key features of industrial data platforms. Our solution has several original key approaches. First of all, its architecture follows the ‘privacy-by-design’ principles by developing the platform’s functionalities using our own blockchain solution centered on data confidentiality mechanisms. Second of all, aspects regarding data reusability, traceability and findability are approached by developing in DataShareFair a set of methodologies we proposed, based on the FAIR principles. Third of all, formalization of legal agreements is possible through direct and group data sharing based on dataset specific access policies defined in the platform. Data governance mechanisms are built on top of the FAIR compliant metadata modelling and our own blockchain solution. They are included in the DataShareFair platform, whose decentralized architecture ensures businesses have full control of the terms and conditions under which their data is used.

2. Data Modelling in B2B Processes

Digital data accessibility and timely sharing have proven to be key aspects in the innovation process of any research area. In Wilkinson (2016), the authors outline the importance of improving the currently used digital ecosystem in order to support automatic reuse of scholarly data. Since there are many research areas that could benefit from this initiative, such as academia, industry, funding agencies and scholarly publishers, FAIR data principles have been established.

2.1. FAIR Principles Overview

The first formal steps towards better data management and governance were made when the “FAIR Guiding Principles for Scientific Data Management and Stewardship” [11] paper was published. Its authors proposed some written guidelines to improve findability, accessibility, interoperability, and reuse of digital assets [12]. These principles can be used to automate the process of extracting metadata from available data. This is a very important aspect, considering the increase in complexity and volume of the information collected these days. According to [13], the European Commission’s Open Research Data Pilot [14] tried to incorporate these principles only a few months later after their publication. Their goal is to make research results more visible in the community and easier to share between institutions. However, extensive research still needs to be performed until the FAIRification process can be applied successfully to the variety of open data available at this point on the Web [14].

The main rules proposed in the FAIR data principles paper should be used in the process of FAIRification, whose purpose is to make existing data compliant with the FAIR model.

The first characteristic of a dataset should be findability. To achieve this, metadata extracted from data and provided by the authors should be rich in terms of volume and references to already standardized concepts. As an additional measure to increase its findability on the long term, a unique and persistent identifier should be provided for the shared data.

Accessibility is another key requirement when sharing and reusing the published data, according to the FAIR model. Metadata and data should be easily processed by both machines and humans, while being made accessible for manual and automatic discovery on a repository. FAIR data does not necessarily mean Open Data, which suggests that the data repository and its contents may be made public under certain sharing agreement licenses, while also requiring the user to authenticate and receive an authorization to

access the data. Following the formerly mentioned principles when implementing a data management system also contributes to the data governance process [15,16], one of the main issues in the case of Open Data.

Since one of the main goals of sharing data is integrating it with existing datasets used in processing workflows, interoperability appears to be another important property. The constructed metadata should be formally defined using an accessible and broadly applicable knowledge representation language, in order to be uniform in structure and easily integrateable in existing processing and storage tools.

The main purpose of the FAIR methodology is to “optimize data reuse”. Combination of data in different contexts can only be achieved if the data and infrastructure are described by rich metadata. All data collections should provide clear information about usage licenses in order to enable references in the new metadata. The data governance process is facilitated by having accurate information about the provenance and distribution license.

2.2. Open Data Debates Approached in the FAIR Data Model

All automatic processes have one common thing they need in order to produce relevant results—an abundance of data from various sources. When it comes to improving the current living standards of our society, the most important data publishers are public sector actors, such as governmental institutions and NGOs. For the European Union member countries, the data sharing and disclosure process is subjected to regulations imposed both at national and communitarian level.

Modelling the information produced by the public representatives in agreement to the open standards is part of the European Data Portal Initiative [17], which intends to make the data freely available and accessible for reuse on any purpose. Even though this initiative is partially regulated by the “Public Sector Information Directive” [18] and it brings a lot of benefits, such as free “flow of data, transparency and fair competition” [17], there are several legal and technical issues that arise from the disclosure and uncontrolled usage.

As stated in [19] the first issue that arises in the context of open data implementation and usage is governance [20]. Data supply, access and ownership procedures are not clearly stated, which leads to creation of ad-hoc requirements established in custom negotiations between parties. This behavior overrules the basic characteristics of open data.

Even though the main goal of open data is to be accessible by all, sensitive information collected by governmental institutions should be kept private and disclosed only with the user consent according to GDPR regulations [21]. Using open data to increase reusability of existing information and to fructify research opportunities does not guarantee the user with his entitled privacy, therefore, it raises issues of trust and creates legal barriers.

In terms of FAIR data, these issues can be partially solved by first following the Reusability guidelines, where detailed provenance of the data and metadata must be supplied. Metadata should only be released under a clear usage license. Secondly, the accessibility principles should be followed, since a protocol with support for authentication and authorization is needed to control which entities have access to data and under which circumstances. Availability of data is ensured by adhering to the findability and accessibility rules proposed in the FAIR paper [11]. Effective data governance cannot be applied to open data. There is no way to prove its consistency and trustworthiness, while also tracking its usage.

Another issue that arises when trying to create open data refers to the lack of standardization amongst existing datasets [19]. This leads to poor quality data that cannot be easily associated with other data in automated processes, since labels and time indicators are not used in a consistent manner. Open data guarantee us neither interoperability between old and new, nor provenance of data since there are no written requirements for the shared metadata. This aspect is approached by the FAIR methodology through its interoperability and reusability principles which clearly state the metadata categories and how they should be modelled in order to reach a processing ready dataset.

According to [22], finding and accessing open data is another issue that has not been thought through. The authors emphasize the need to easily locate the data, while also keeping track of the time data were made available and last updated. To achieve these goals, they propose creating open data registries structured relative to location, which should contain datasets identified by various keywords. This solution is already similarly modelled in the FAIR Data Point specification [23]. A FAIR Data Point represents a registry of FAIR datasets and their metadata served through a REST API, among which we can also find information about the required time variables and keywords [23]. The main purpose of a FAIR Data Point is to ensure data discovery, access, publication and metrics.

Several issues arise when it comes to sustaining open data costs of storage, delivery and maintenance. There is no standardized way to keep metadata, while erasing stale data in order to reduce costs. This situation is foreseen by the FAIR principles, which comprise that metadata should not be erased from a FAIR Data Point, even though data is no longer available. The FAIR approach guarantees us a way to easily find historical records of data and its publishers, while also minimizing the costs of keeping perishable data alive.

Data redundancy also appears to be a recurrent problem in the usage of open data. Since there has not been established a clear methodology on how to annotate data in order to be uniquely identifiable, there is no guarantee open means unique. The FAIR data ontology modelling guarantees uniqueness through the reusability principles, which state that each piece of metadata and data has its own identifier.

3. DataShareFair Proposal

3.1. FAIR Data Modelling in DataShareFair

The reference metadata modelling [24] proposed by the FAIR Data team organizes business metadata in a structured way, using entities such as catalogs, datasets and distributions. In this section, we present a FAIR based ontology proposal, used for structuring metadata we extract from datasets provided by business entities. The resulting structured information is managed in our proposed platform, DataShareFair, according to a series of procedures we proposed. Additionally, we need to integrate an access control model in order to respect restrictions that may appear from the licensing terms.

In Table 1, we present the metadata structuring for a Catalog in the context of DataShareFair. The model's properties and meaning were inspired from the FAIR Data Team proposal [24] and adapted to fit the needs of our decentralized data marketplace platform.

If a business representative wants to share data, that person must create or choose a catalog where the particular metadata will be included as a new dataset. A catalog is composed of several datasets which share the same theme taxonomy and other specified metadata. Table 2 describes the annotated metadata properties required to be specified for a FAIR compliant dataset, according to the FAIR Data Team proposal [24].

Table 1. Catalog Metadata Entry in DataShareFair.

<http://192.168.1.52:8087/fdp/catalogs/biohazard_catalog>
a <http://www.w3.org/ns/dcat#Catalog>;
<http://www.w3.org/2000/01/rdf-schema#label>
"Biohazard catalog";
<http://purl.org/dc/terms/accessRights>
<http://192.168.1.52:8087/fdp/catalogs/biohazard_catalog#accessRights>;
<http://purl.org/dc/terms/conformsTo>
<https://www.purl.org/fairtools/fdp/schema/0.1/catalogMetadata>;
<http://purl.org/dc/terms/description>
"A catalog that should contain datasets about artificial produced disasters in nature; data can be used to predict future disasters.";
<http://purl.org/dc/terms/hasVersion>

Table 1. Cont.

"1";
 <http://purl.org/dc/terms/identifier>
 "biohazard_catalog";
 <http://purl.org/dc/terms/isPartOf>
 <http://192.168.1.52:8087/fdp/>;
 <http://purl.org/dc/terms/language>
 <http://id.loc.gov/vocabulary/iso639-1/cy>;
 <http://purl.org/dc/terms/license>
 <https://www.gnu.org/licenses/fdl-1.2.html>;
 <http://purl.org/dc/terms/publisher>
 <http://192.167.1.53:8087/fdp/organizations/CC28732>;
 <http://purl.org/dc/terms/title>
 "A new bright catalog";
 <http://rdf.biosemantics.org/ontologies/fdp-o#metadataIdentifier>
 <http://192.168.1.52:8087/fdp/catalogs/biohazard_catalog>;
 <http://rdf.biosemantics.org/ontologies/fdp-o#metadataIssued>
 "1586882492976""<http://www.w3.org/2001/XMLSchema#dateTime>;
 <http://rdf.biosemantics.org/ontologies/fdp-o#metadataModified>
 "1586882492976""<http://www.w3.org/2001/XMLSchema#dateTime>;
 <http://www.w3.org/ns/dcat#dataset>
 <http://192.168.1.52:8087/fdp/datasets/biohazard_dataset>, <http://192.168.1.54:8087/fdp/datasets/natural_disasters>;
 <http://www.w3.org/ns/dcat#themeTaxonomy>
 <http://dbpedia.org/resource/Category:Natural>, <http://dbpedia.org/resource/Category:Biohazard>.

Table 2. Dataset Metadata Entry in DataShareFair.

<http://192.168.1.52:8087/fdp/datasets/biohazard_dataset>
 a <http://www.w3.org/ns/dcat#Dataset>;
 <http://www.w3.org/2000/01/rdf-schema#label>
 "Biohazards Dataset";
 <http://purl.org/dc/terms/accessRights>
 <http://192.168.1.52:8087/fdp/datasets/pretty_dataset#accessRights>;
 <http://purl.org/dc/terms/conformsTo>
 <https://www.purl.org/fairtools/fdp/schema/0.1/datasetMetadata>;
 <http://purl.org/dc/terms/description>
 "Dataset that contains data about different kind of biohazards";
 <http://purl.org/dc/terms/hasVersion>
 "1";
 <http://purl.org/dc/terms/isPartOf>
 <http://192.168.1.52:8087/fdp/catalogs/biohazard_catalog>;
 <http://purl.org/dc/terms/language>
 <http://id.loc.gov/vocabulary/iso639-1/da>;
 <http://purl.org/dc/terms/license>
 <https://creativecommons.org/licenses/by/3.0/>;
 <http://purl.org/dc/terms/publisher>
 <http://192.167.1.53:8087/fdp/organizations/CC28732>;
 <http://purl.org/dc/terms/title>
 "Biohazards Dataset";
 <http://rdf.biosemantics.org/ontologies/fdp-o#metadataIdentifier>
 <http://192.168.1.52:8087/fdp/datasets/biohazard_dataset>;
 <http://rdf.biosemantics.org/ontologies/fdp-o#metadataIssued>
 "2020-04-14T18:23:28.668Z""<http://www.w3.org/2001/XMLSchema#dateTime>;
 <http://rdf.biosemantics.org/ontologies/fdp-o#metadataModified>
 "2020-04-14T18:23:28.668Z""<http://www.w3.org/2001/XMLSchema#dateTime>;
 <http://www.w3.org/ns/dcat#distribution>
 <http://192.168.1.52:8087/fdp/distributions/biohazard_danger_statistics>;
 <http://www.w3.org/ns/dcat#keyword>
 "hazard", "biology", "nature";
 <http://www.w3.org/ns/dcat#theme>
 <http://dbpedia.org/resource/Category:Hazard>, <http://dbpedia.org/resource/Category:Nature>.

After a business entity defines a dataset and its metadata, several data distributions can be attached to it. A distribution is linked to the provided data by the publisher through the intrinsic extracted metadata. Creating a new distribution and attaching it to a dataset is equivalent to making a new version of the previously shared data available.

As an overview, one catalog contains a number of datasets, which are composed of distributions containing direct links to the data bundles. Table 3 describes the annotated metadata properties required for a distribution to be compliant with the FAIR methodology, according to the FAIR data team proposal [24].

Table 3. Distribution Metadata Entry in DataShareFair.

```

<http://192.168.1.52:8087/fdp/distributions/biohazard_danger_statistics>
  a    <http://www.w3.org/ns/dcat#Distribution>;
  <http://www.w3.org/2000/01/rdf-schema#label>
    "Biohazard Dangers Statistics";
  <http://purl.org/dc/terms/accessRights>
<http://192.168.1.52:8087/fdp/distributions/biohazard_danger_statistics#accessRights>;
  <http://purl.org/dc/terms/conformsTo>
<https://www.purl.org/fairtools/fdp/schema/0.1/distributionMetadata>;
  <http://purl.org/dc/terms/description>
    "This distribution aims to offer data about existing biohazardous substances that can be found and in which quantities";
  <http://purl.org/dc/terms/hasVersion>
    "1";
  <http://purl.org/dc/terms/isPartOf>
<http://192.168.1.52:8087/fdp/datasets/biohazard_dataset>;
  <http://purl.org/dc/terms/language>
<http://id.loc.gov/vocabulary/iso639-1/ch>;
  <http://purl.org/dc/terms/license>
<https://www.apache.org/licenses/LICENSE-1.0>;
  <http://purl.org/dc/terms/publisher>
<http://192.167.1.53:8087/fdp/organizations/CC28732>;
  <http://purl.org/dc/terms/title>
    "Biohazard Dangers Statistics";
  <http://rdf.biosemantics.org/ontologies/fdp-o#metadataIdentifier>
<http://192.168.1.52:8087/fdp/distributions/biohazard_danger_statistics>;
  <http://rdf.biosemantics.org/ontologies/fdp-o#metadataIssued>
    "2020-04-14T18:26:35.097Z"^^<http://www.w3.org/2001/XMLSchema#dateTime>;
  <http://rdf.biosemantics.org/ontologies/fdp-o#metadataModified>
    "2020-04-14T18:26:35.097Z"^^<http://www.w3.org/2001/XMLSchema#dateTime>;
  <http://www.w3.org/ns/dcat#downloadURL>
<http://192.168.1.52:8087/uploads/b4db06ef-ce6e-4005-976a-e93385dad18a.zip>;
  <http://www.w3.org/ns/dcat#mediaType>
    "application/gzip".

```

Each created catalog, dataset and distribution resource has an access policy IRI assigned through the "dct:accessRights" property, present in each of the Tables 1–3. A new authorization granted to an organization or entity over a resource—either read, write or control rights—is linked to its access policy. The access policy model used for the FAIR modelled resources is based on the Web Access Control Vocabulary (WAC) [25].

Any new authorization created in DataShareFair will follow the modelling described in Table 4, by using a part of its properties. Each authorization will be linked to the corresponding access policy of the catalog, dataset or distribution.

Another concept needed in order to model authorizations and also include information about the publisher of the data is the organization. The proposed organization modelling can be viewed in Table 5.

Table 4. Access Policy and associated Authorizations Metadata Entries in DataShareFair.

http://192.168.1.52:8087/fdp/distributions/cat_sweet_dist#accessRights	
a	http://purl.org/dc/terms/RightsStatement ;
	http://purl.org/dc/terms/description
	"This resource has access restriction";
	http://purl.org/dc/terms/isPartOf
	http://192.168.1.52:8087/fdp/authorizations/9f4603eb-120b-45be-9fbf-c2f533d7b2bb ,
	http://192.168.1.52:8087/fdp/authorizations/1f3b1276-2d10-436f-8b38-55655b1c73c3 .
http://192.168.1.52:8087/fdp/authorizations/9f4603eb-120b-45be-9fbf-c2f533d7b2bb	
a	http://www.w3.org/ns/auth/acl#Authorization ;
	http://www.w3.org/ns/auth/acl#agent
	http://192.165.1.54:8087/fdp/organizations/CC29751 ;
	http://www.w3.org/ns/auth/acl#mode
	http://www.w3.org/ns/auth/acl#Read .
http://192.168.1.52:8087/fdp/authorizations/1f3b1276-2d10-436f-8b38-55655b1c73c3	
a	http://www.w3.org/ns/auth/acl#Authorization ;
	http://www.w3.org/ns/auth/acl#agent
	http://192.167.1.53:8087/fdp/organizations/CC28732 ;
	http://www.w3.org/ns/auth/acl#mode
	http://www.w3.org/ns/auth/acl#Control .

Table 5. Organization Metadata Entry in DataShareFair.

//Business Organization URI type specification	
http://192.165.1.54:8087/fdp/organizations/CC29751	
a	http://xmlns.com/foaf/0.1#Organization ;
//property which specifies the unique identifier of the current resource	
http://schema.org/Thing#identifier	
http://192.165.1.54:8087/fdp/organizations/CC29751 ;	
//property which specifies the official web address of a business—for additional information	
http://schema.org/Thing#url https://bio-research.com ;	
//property which specifies the popular/known name of the business	
http://xmlns.com/foaf/0.1#name	
"Biohazard Research Group" ;	
//property which specifies an email contact address of the business	
https://schema.org/Organization#email contact@biohazard-research.com	
//property which specifies the legal and official name of the business	
https://schema.org/Organization#legalName "Biohazard Research Group" ;	
//property which specifies the fiscal identifier for a business—also used to compose the unique identifier of the business	
https://schema.org/Organization#vatID "CC29751" ;	
//property which specifies the official residence for a business—either the full address or just the global region	
https://schema.org/Organization#address "Europe" .	

Using the hierarchical data structuring in catalogs, datasets and distributions, while modelling the metadata accordingly to the proposed models is one of the first steps in providing FAIR data in terms of interoperability and reusability. In order to respect enterprise data governance requirements in the sharing process, modelled metadata was used in several mechanisms and processes, presented in detail in Section 3.2.

In Table 6, we argue that the metadata modelling process incorporated in DataShare-Fair is compliant with the FAIR principles.

Table 6. FAIR principles applied in DataShareFair data and metadata modelling.

Category	Main Principle	Subprinciples	DataShareFair Data Modelling Applicability
Findable	F1. (Meta)data are assigned a globally unique and persistent identifier	—	Each dataset, catalog, distribution and organization resource has a uniquely generated identifier (<i>fdp:metadataIdentifier</i> property in Tables 1–3 and <i>thing:identifier</i> property in Table 5)
	F2. Data are described with rich metadata (defined by R1 below)	—	Each distribution resource contains information about the issued and last update dates, keywords and intrinsic metadata that can be extracted from the uploaded distributions (<i>fdp:metadataIssued</i> and <i>fdp:metadataModified</i> from Table 3); this allows easy data discovery by automated processes or human filtering
	F3. Metadata clearly and explicitly include the identifier of the data they describe	—	Each Dataset contains a list of IRIs for all the available distributions (<i>dcat:distribution</i> from Table 2); distribution metadata is linked to the dataset metadata through the <i>dcat:dataset</i> property from Table 3; dataset metadata is linked to the catalog metadata through the <i>dct:isPartOf</i> property from Table 2
	F4. (Meta)data are registered or indexed in a searchable resource	—	The existing dataset metadata and its distribution's metadata is indexed for human search in the DataShareFair Management Platform; for automated search, data will be available at FAIR data points in the network; this process is presented in Section 3.2
Accessible	A1. (Meta)data are retrievable by their identifier using a standardized communications protocol	A1.1 The protocol is open, free, and universally implementable A1.2 The protocol allows for an authentication and authorization procedure, where necessary	The used protocol for exchanging metadata in DataShareFair is swarm communication [26], which is open, free and universally implementable; exchanging data through the distributed fair data points can be achieved through the HTTPs communication; communication protocols are detailed in Section 3.2. Each business representative is authenticated with its own credentials to access shared data; for each request an authorization token is provided to validate the requestor's identity
	A2. Metadata are accessible, even when the data are no longer available	—	Stale data deletion is allowed; metadata remains intact and cannot be deleted; A catalog of datasets can never be deleted, even though the requester is one of the owners; this restriction is controlled from our FAIR data management platform, presented in Section 3.2.
	I1. (Meta)data uses a formal, accessible, shared, and broadly applicable language for knowledge representation	—	Catalogs, datasets, distributions and other resources will be stored in a graph-oriented database (Apache Jena) and they will be available in several interoperable and machine-readable formats such as: JSON-LD, OWL, XML; additional details are provided in Section 3.2.1
Interoperable	I2. (Meta)data use vocabularies that follow FAIR principles	—	The vocabularies used to describe the datasets, catalogs, distributions are standardized and have external IRIs open for access to anyone interested; examples of such vocabularies: DCTERMS, FDP, RDF, XSD, DCAT—used in Tables 1–5
	I3. (Meta)data include qualified references to other (meta)data	—	Each dataset resource contains a list of IRIs for all the available distributions using DCAT ontology (<i>dcat:distribution</i> property in Table 2)

Table 6. Cont.

Category	Main Principle	Subprinciples	DataShareFair Data Modelling Applicability
Reusable	R1. Meta(data) are richly described with a plurality of accurate and relevant attributes	R1.1. (Meta)data are released with a clear and accessible data usage licens	The type of license under which the data are distributed is mentioned in each Catalog, Dataset and Distribution resource (<i>dcterms:license</i> property in Tables 1–3)
		R1.2. (Meta)data are associated with detailed provenance	This principle is covered by the <i>dcterms:license</i> and the <i>dcterms:publisher</i> properties in the distribution modelling from Table 3; Each distribution instance is connected to several access rights instances—in Table 4; they describe the limits of usage for a dataset reported to organizations—described in Table 5
		R1.3. (Meta)data meet domain-relevant community standards	Data shared by the businesses is organized in a standardized way by extracting and structuring the metadata according to the ontology specification; The storage of the distributions is done in a well-established and sustainable file format; Documentation (metadata) follows a common template and uses common vocabulary

Table 6 shows how the approached modelling of the metadata promises to make the data findable, accessible, interoperable and reusable. The FAIR principles and subprinciples are underlined relative to the metadata modelling of the targeted resources, presented in Tables 1–5 and their usage. The management platform for FAIR data, which we propose in this paper, applies several processes over the shared data, using the proposed metadata structuring. These processes are needed in order to ensure businesses can assert the ownership of their data, control its usage and easily integrate it, while preserving its compliance with the data confidentiality legislation.

3.2. DataShareFair Technical Approach

The emerging trend of including B2B data sharing in various business strategies and its advantages were highlighted in Section 1. The main issues in this process consist in different data formats and volumes, the need for technical customization, broken trust and legal barriers and the lack of ownership assertion and control of use over the data. We propose DataShareFair as a viable solution for intra- and cross-domain/sector FAIR data sharing which addresses these problems. Our focus is to fulfill diverse privacy and security [27,28] needs which arise in an enterprise context. We also aim to facilitate the creation of efficient business models.

3.2.1. Underlying Concepts and Technologies

The DataShareFair solution was developed using known technologies and frameworks, presented in Figure 1. The resource management Web interface offered to business representatives to access the platform functionalities was developed using the angular framework [29]. The FAIR metadata is stored and manipulated using the Apache Jena triple store database [30], a free and open-source Java framework for building semantic web and linked data applications. Since data and metadata need to be accessible through a Fair data point integration, a Jena Fuseki server [31] was deployed in order to expose information through a HTTPs endpoint. Packaged delivery of all components included in the proposed solution is possible through Docker technology [32]. They all run in a single container, exposed through a configured virtual network, in order to be accessed by other nodes.

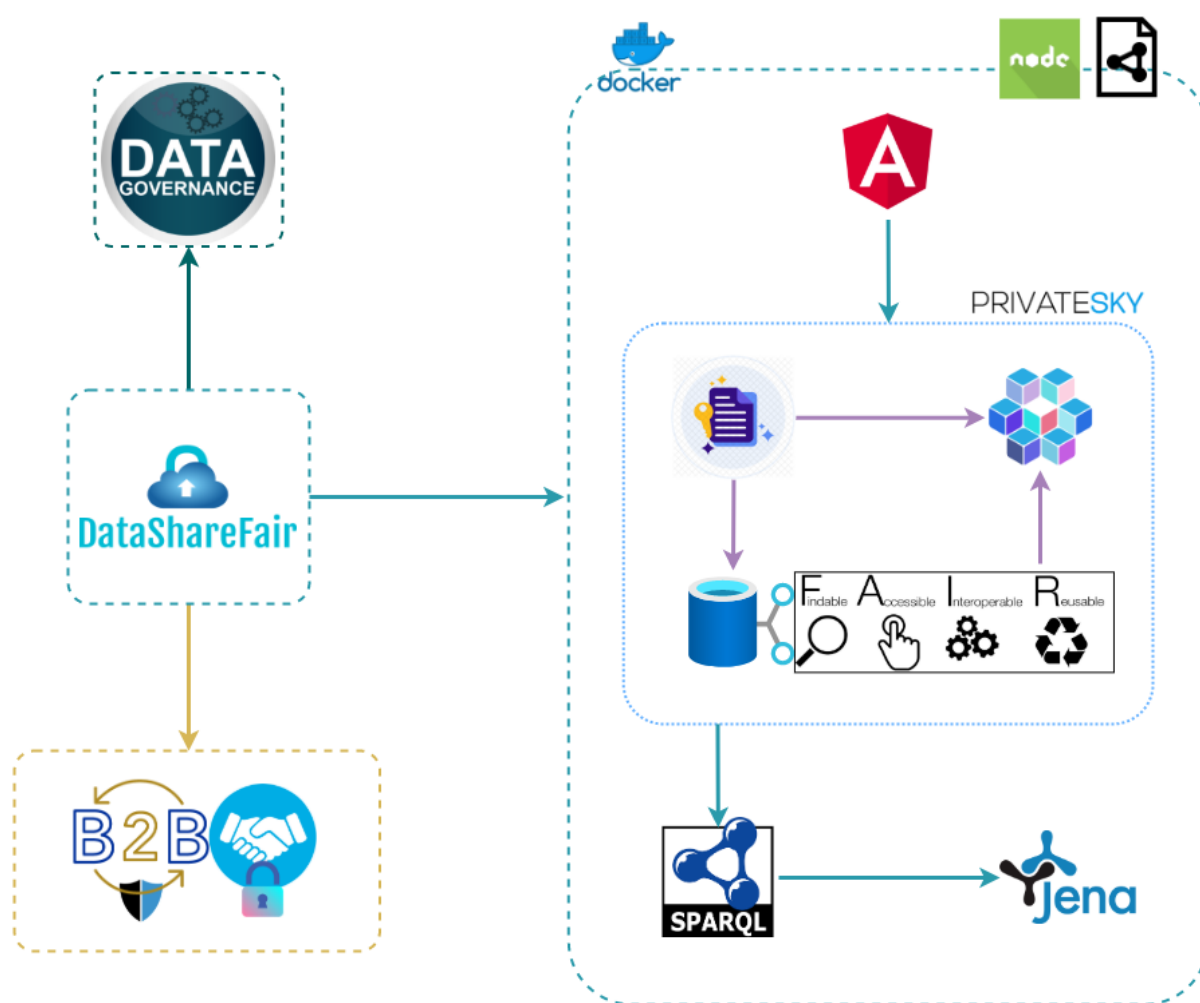


Figure 1. DataShareFair Architecture Overview.

The underlying functionalities of the platform in terms of data FAIRification and management were built using our own version of smart contracts and assets over blockchain. We contributed to these technologies as collaborators in the PrivateSky European Project [25]. Its first version, the SwarmESB project [33], proposed a lightweight enterprise service bus based on the swarm communication pattern. It enabled the build of scalable enterprise systems, based on multiple decoupled components. In order to ensure data confidentiality mechanisms are integrated in PrivateSky, the SwarmESB core architecture was adapted to facilitate the definition and usage of smart contracts and their interaction with blockchain based technologies. A comprehensive example of a system developed in the Swarm ESB ecosystem, which targets the need for data confidentiality and user privacy was described in [21]. This system showed how public institutions could adapt their procedures in order to comply with the legal prerequisites imposed by the GDPR principles. The experience we obtained from modelling its architecture in respect to the “privacy by design” principles [34] heavily contributed to our current proposal.

PrivateSky’s open source platform is based on SwarmESB and offers the possibility to create several integration layers for cloud and enterprise applications. The proposed architecture focuses on microservices and transactions. Their composability is accomplished by using swarm communication. PrivateSky platform enables the development of decentralized applications, relying on blockchain technology. It addresses security and privacy needs through consortium and private distributed ledgers [35]. The main subjects of innovation approached by this platform are swarm communication, blockchain storage and handling of private data using EDFS [36].

Swarm messaging represents a communication pattern, which allows sending and processing messages among different modules of a software system. These specialized messages are capable of taking on responsibilities in the orchestration process of the components defined in the PrivateSky ecosystem. A set of such connected messages defines a swarm. When executed, each swarm launches one or many flows during its processing. On processing completion, the computational resources are freed, according to their definition in respect to the serverless paradigm [37]. Choreographic smart contracts are a type of swarms, which focus on operations that need to be performed on the blockchain. This is possible by defining a collection of assets and transactions in order to build a workflow manageable under the consensus and regulations of a blockchain. A smart contract can be launched into execution on a network node only if a security context, containing secret data and actor identities, is used to authorize operations [37]. A blockchain asset represents a serializable object which can have different states (versions) at a point in time. A transaction is a flow of processing which transitions an asset from State A to State B.

PrivateSky does not propose a single monolithic blockchain but a set of hierarchical blockchains called “blockchain domains” [27], as it can be seen in Figure 2. As a result, the smart contracts built on top of these resources are stored and executed inside of a CSB Constitution (Cloud Safe Box). This way they are considered secret from the anchoring domain point of view. This approach solves privacy related problems without reducing security, since the number of nodes in the network can be kept big for consensus operations without sharing unnecessary information about the input, output and the processing performed on data in the smart contract [38].



Figure 2. Multichain Blockchain Logical Domains Architecture [39].

PrivateSky’s blockchain architecture is based on multiple tiers, as it is presented in Figure 3. The “Networking” layer sets the rules for communication between nodes, starting from booting to validating access restrictions by certificates. The “Blockchain Logical View” layer is described by the conceptual multichain domains from Figure 2. The “Physical Blockchain Replicas and Storage” level refers to the actual software running on nodes in order to participate in consensus, caching or encrypted storage operations that are performed for transactions to be anchored. The “Containerized APIs” level exposes services containing logic to perform transactions on blockchain in the form of REST APIs or

smart contracts, depending on each application's purpose and goal. The "Far-Chain: Block Applications" level includes applications which invoke through swarm communication [40] the functionalities encapsulated in the smart contracts from the previously described architectural layer.

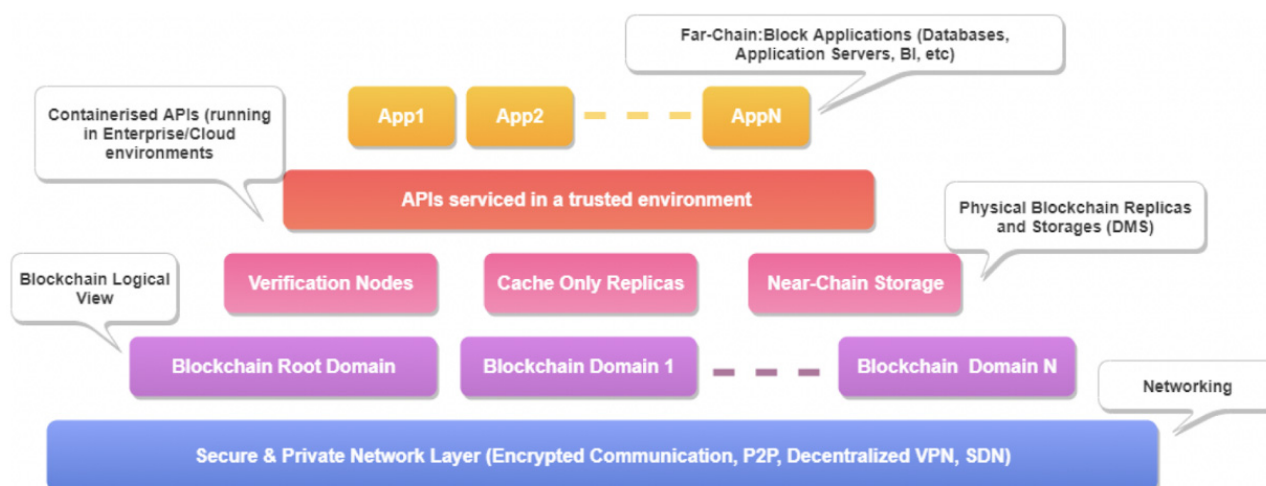


Figure 3. Blockchain Application General Architecture [39].

From a practical view, one of the blockchain domains from Figure 2 can be seen as a chain of immutable blocks. They are created as a result of the "Far-Chain: Block Applications" level in Figure 3 requesting services from the "Containerized APIs" level. These services are offered through smart contracts, which gather sets of transactions over the blockchain. In order to add a new block, the set of transactions proposed by a node must be accepted in consensus by all the other nodes in the network. Consensus in the PrivateSky's blockchain is performed using a custom algorithm, designed for replicated distributed systems which require Byzantine fault tolerance. It was developed as part of the PrivateSky project and its efficiency was validated through a specific testing strategy. A main characteristic for this algorithm is the fact that it uses digital signatures and broadcast to record network activity, which is a different approach to the classical blockchain architecture. The consensus participants inform everyone else about the set of accepted or rejected transactions in a time frame.

3.2.2. Proposed Architecture

The designed architecture is presented in Figure 4. Our platform is fully decentralized, since each consortium node runs its own instance of the DataShareFair system, composed of two components. This allows each node in the network to submit transactions on the FAIR metadata model and vote for their persistence over the blockchain. Each network node is independent of its peers in the process of making data and metadata available in the system according to self-defined access policies, which allows better control over the shared data. Due to its P2P architectural model, each node can listen to multiple connections from other network nodes. On creation of a successful connection, data can be directly transferred between the nodes, without the need of a central authority to intermediate the process. In terms of node failure, the node stored data and metadata may no longer be available to the other network members until recovery. Fortunately, failure of one node causes only part of the data to be temporarily unavailable, while the system is not affected as a whole.

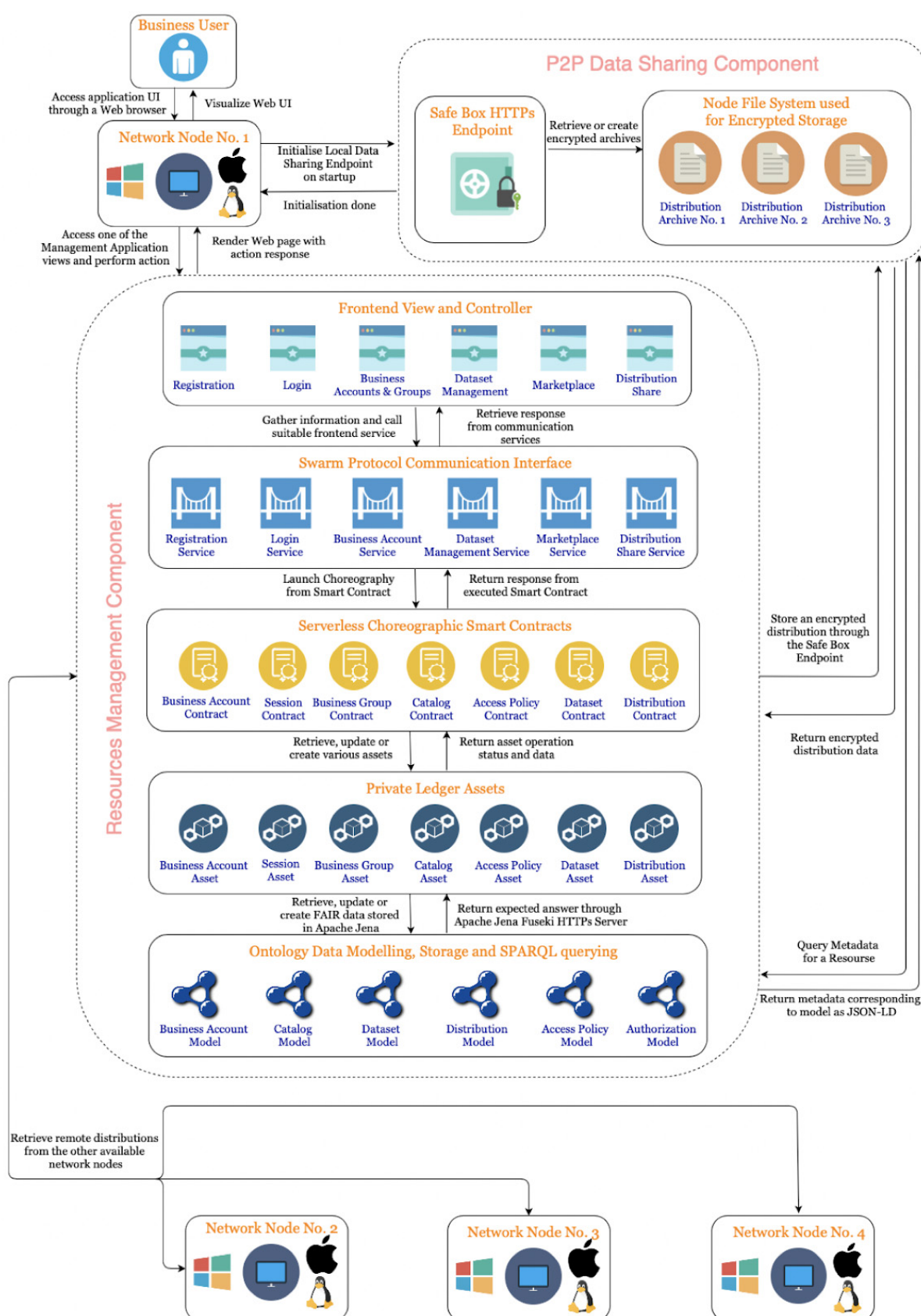


Figure 4. DataShareFair Platform Architecture.

The main component of our proposed architecture is the “Resources Management” module, which is run on each node to obtain access to the data sharing and management features. After the business representative provides proof of its identity through a digital certificate, it is able to find and manage FAIR data directly on its own node through a locally hosted Web interface, described in the “Frontend View and Controller” layer from Figure 4. Each operation executed in the Web interface is passed to the next layer, the Swarm Protocol Communication Interface. In this layer, the provided data is packaged and the necessary smart contract is run, depending on the action performed in the interface. The third architectural layer includes all smart contracts, which proved to be highly efficient in terms of previously developed decentralized applications [41]. The available smart contracts encapsulate the actual business logic to manage and exchange data between parties. PrivateSky smart contracts are easily maintainable, since they can be updated and versioned based on flow changes. Even though their execution is started on the local node, the resulting transactions over the assets in the fourth level are broadcasted to all the P2P network nodes. These transactions are validated by the network nodes using a consensus algorithm, which upon success commits them in the underlying blockchain. The used blockchain has the role to keep references towards the semantic models of catalogs, datasets and distributions metadata stored on the node. The references are created as assets in the blockchain and their information and version are updated with each accepted transaction, as it is outlined in the fourth layer of the component. Each node saves the transactions in its own ledger and assets are synchronized to have the latest value. The semantic FAIR metadata represented in the fifth layer is updated when transactions are accepted over the assets, which guarantees data integrity at all times. The mechanisms used to validate transactions allow us to keep track of the executed changes through the immutable blocks stored in the blockchain. Access to data offered by nodes in the network can be provided only through the management interface. Since gaining access to any type of data requires authentication in the Web platform and an authorization level provided by the owner, we address several known security and privacy issues of legal and technical nature identified in other B2B data exchange platforms.

The second module in our solution, outlined in Figure 4, is the P2P data sharing component. Its name emerges from the direct communication between nodes, which is performed in the actual data exchange process. Therefore, as it is specific to a fully decentralized system, no central authority and no intermediaries are needed in order to locate and identify the data, since this information is in each node’s ledger. The main purpose of this module is to store and allow the retrieval of FAIR data and metadata by validating if requesting nodes are authorized to perform the action. When a business decides to create a new distribution and attach it to a dataset, the distribution’s data is stored encrypted on the node’s file system, in order to prevent its unauthorized usage. Access to this data is provided only through the management platform, which runs a smart contract to decrypt the contents before making them accessible. Another purpose of this component is to provide access to a node specific FAIR Data Point. More explicitly, each node from the P2P network overlay exposes its own FAIR data and metadata in an independent and controllable manner, which is specific for decentralized systems. In comparison to the traditionally proposed FDP [23], our solution allows a business to assert ownership and control of use over its data through encryption, reversible only on validation of the defined access policies.

Building our system in accordance with a P2P decentralized architecture model solves a series of issues which are commonly encountered in centralized systems. First of all, performance bottlenecks are avoided by allowing each node to run its own instance of DataShareFair. In our scenario, each node in the network will be responsible to process its own operation and synchronize its state with the other members. Nonetheless, in a centralized system, a server containing all the functionalities encapsulated in the smart contracts would have to serve all the clients and perform changes over the central semantic model, which would clearly result in a bottleneck on high request rates. Second of all,

our system has high availability, since it is only partially affected from the data source point of view when a node fails. Third of all, each node has autonomy and control over its shared resources, since it can offer or deny access to data based on access policies. These are defined locally and shared through the network by synchronizing the approved transactions among each node's ledger.

DataShareFair encapsulates the proposed B2B data sharing methodologies and other techniques to offer businesses governance over their data. The main functionalities of the proposed solution [42], available for a business representative, are structured as follows:

1. Registration and Login

Any business representative can take advantage of the DataShareFair capabilities together with their business partners by launching the Docker image of the system. This image should run in a container residing on a node included in an enterprise permissioned network formed by these businesses. The resource management module is accessible locally through a web client. By default, the business representative will be able to log in using their VAT or TIN unique identifier [43] and their digital company certificate. The provided information will be gathered in the "Login" view from the "Frontend and Controller" layer of the multi-tiered component from the architecture in Figure 4. This information will be used by the "Login" angular service to launch into execution various processing phases from the "Session" smart contract mentioned in the third layer of the Resources Management component from Figure 4. This smart contract will record a new entry for the "Session" asset defined in our blockchain. The logout functionality invalidates an instance of a "Session" asset by updating its expiration status.

There is also a registration feature provided for a new business. The interaction flow between components from each tier is similar to what we presented in the login flow. As a difference, the business entity accesses the "Registration" view and the information will be formatted as a swarm message in the "Registration" service. The registration phase implemented in the "Business" smart contract is executed and this results in the creation of a new "Business" asset instance. A new entry in the "Business" collection from our Triplestore is created and formatted according to the semantic modelling presented in Table 5.

2. Business Accounts and Groups Management

Business representatives can update some of the information initially provided at registration through the "Business Account" service included in the second layer of the architecture. Some specific phases will be invoked through swarm communication from the "Business Account" smart contract in the third layer. These values will be updated in the blockchain by creating a new version of the targeted "Business Account" asset. Another option available is to delete the account, which results in adding empty values in the asset fields and committing a new transaction with them, while also invalidating all active sessions. The semantically modelled metadata stored about the "Business Account", as presented in Table 5, will be changed only after the commit performed in the blockchain is successful.

The second section from Figure 5 allows a business entity to create collaboration groups, which can be used to grant an access level for a metadata resource catalog, dataset or distribution to several members at once. The owner of the group can add any other business entity using the platform in the same network and can revoke access when needed. All these functionalities are available in the "Business Account" service, which runs several phases from the "Business Group" smart contract. Changes on groups are directly performed on versioned assets updated in the blockchain.

Figure 5. Business Accounts and Groups Management View (First Architectural Layer).

3. B2B FAIR Data Marketplace

In Figure 6, we present the data marketplace view. The menu available in the upper left corner allows switching between all the platform functionalities. The data marketplace view offers a business entity an overview over the datasets each business partner has to offer. These datasets can be filtered by using input keywords in the search bar or by selecting one or multiple catalogs from the ones available. The option to search data based on its category is also provided, considering that categories are given when the dataset is created or updated. Dataset categories reference existing concepts described in semantic data sources, such as DBpedia [44] and Wikidata [45]. By default, access to the data distributions of each dataset is restricted. A business representative can request an access level on it, specifically read, write or control. A notification will be sent towards the owners, visible in the menu from the upper right corner of Figure 6. After receiving access to a distribution, the business representative will download it through the Fair Data Point of its business partner, encapsulated in the P2P Data Sharing Component. Decryption of the distribution will be initiated from the client side and performed in the “Distribution” Smart Contract, which first validates the company’s access rights. All these functionalities are accessible through the “Marketplace” service in the second architectural layer from Figure 4. This service will run multiple phases from the “Session”, “Dataset”, “Access Policy”, “Catalog” and “Distribution” smart contracts. The corresponding assets are queried and semantic metadata about datasets, access policies, catalogs and distributions is retrieved from the Triplestore in the fifth layer of the resources management component in Figure 4.

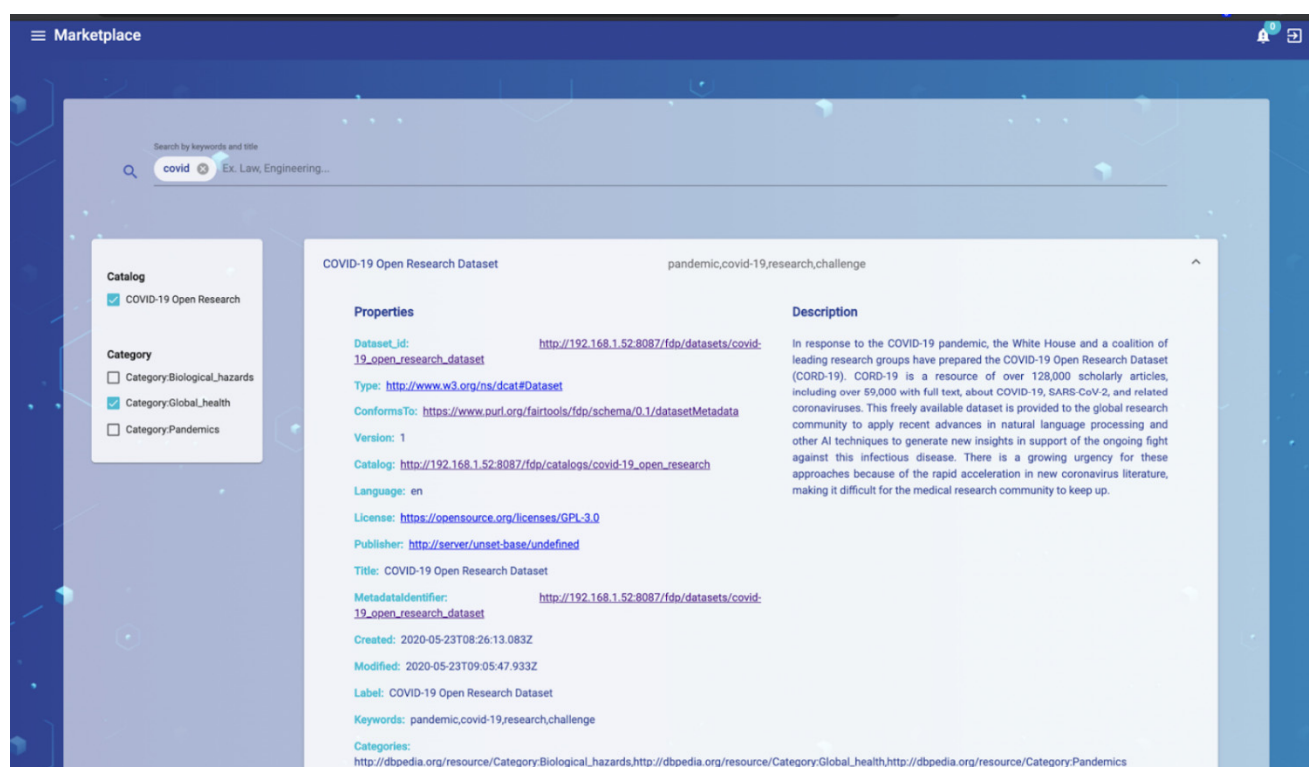


Figure 6. FAIR Data Marketplace Dataset View (First Architectural Layer).

4. FAIRification Process for B2B Data Sharing

Sharing data between multiple businesses is possible through the distributions listed in the marketplace in association with the available datasets. The process of creating a distribution is illustrated in Figure 7, where the business representative will add its name, a short description, the distribution metadata language and the licensing model under which it is shared. The distribution should be included in an existing catalog and dataset or a new instance of each can be created, as shown in the right side of the form in Figure 7. The collected information is used to semantically model the catalog, dataset and distribution, as presented in Tables 1–3. Access rights on the created distribution can be set by adding several partner businesses in the list with an access level. All the access rights granted on a distribution, catalog or dataset will be modelled under an access policy, as shown in Table 4. In the “Distribution Upload” section, we can upload multiple files in various formats. These files will be bundled as an encrypted archive using the ‘AES-256-CTR’ cipher, representing the distribution data to be shared. The created archive will be forwarded towards the SafeBox endpoint, where it will be stored locally on the node’s file system. The decryption key is stored as a private property of the distribution asset, in the blockchain. The feature presented in this view is accessible to the business through the “Distribution Share” service, which uses swarm communication to invoke several phases from the “Distribution”, “Catalog”, “Datasets” and “Access Policy” smart contracts. These smart contracts will create new “Distribution”, “Catalog”, “Datasets” and “Access Policy” assets, included in the set of transactions to be validated by the business partners in the network. Afterwards, the corresponding semantic models are followed to create new entries in the Triplestore.

Distribution

Distribution Metadata

Distribution Title *
Heart Disease UCI

Distribution Description *
This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particu

Distribution Version *
1

Distribution Language
English

Distribution License
Apache Software License

Dataset
Heart Disease UCI Dataset

Catalog
Heart Disease UCI Catalog

ADD DATASET

ADD CATALOG

Distribution Access Rights

Entity Name	Entity ID	Access Level	Actions
SAGETECH MEDICAL EQUIPMENT LIMITED UPDATED	http://192.168.100.15:8087/fdp/GB226345811	Control	
Medical Research Group	http://192.168.100.15:8087/fdp/Medical_Research_Group_YN5edEze1vbgXspcH8uqKUnWfO9nMmElkj09frbvGm0Read		

Add Entity By
Identifier (VAT/TIN_ID or Group ID)

Search String *

Access Right
Read

Add a new member

Distribution Upload

Upload queue

Queue Length: 3

Name	Size	File Type	Actions
00039b94e6cb7609ecbdee1755314bcfeb77faa.json	0.047MB	application/json	
0003793cf9e709bc2b9d0c8111186f78fb73fc04.json	0.023MB	application/json	
0001418189999fea7f7cbe3e82703d71c85a6fe5.json	0.031MB	application/json	

Queue Progress

Remove all

Submit Distribution

Figure 7. FAIR Distribution Creation View (First Architectural Layer).

5. FAIR Resources Management

The metadata management view shown in Figure 8 allows a business owner to visualize all its catalogs, datasets and distributions and those which were shared with it. Catalogs, datasets and distributions can be edited by a partner business, if the owner grants the partner write access to the resources. Distribution bundles can also be deleted if the business has control access. Deleting a distribution bundle does not imply that the associated metadata will be deleted, but rather that the encrypted archive will no longer be stored on the owner business node for download. This action is performed through the Safebox Endpoint, after certain access rights and session permissions are validated in the “Distribution” smart contract.

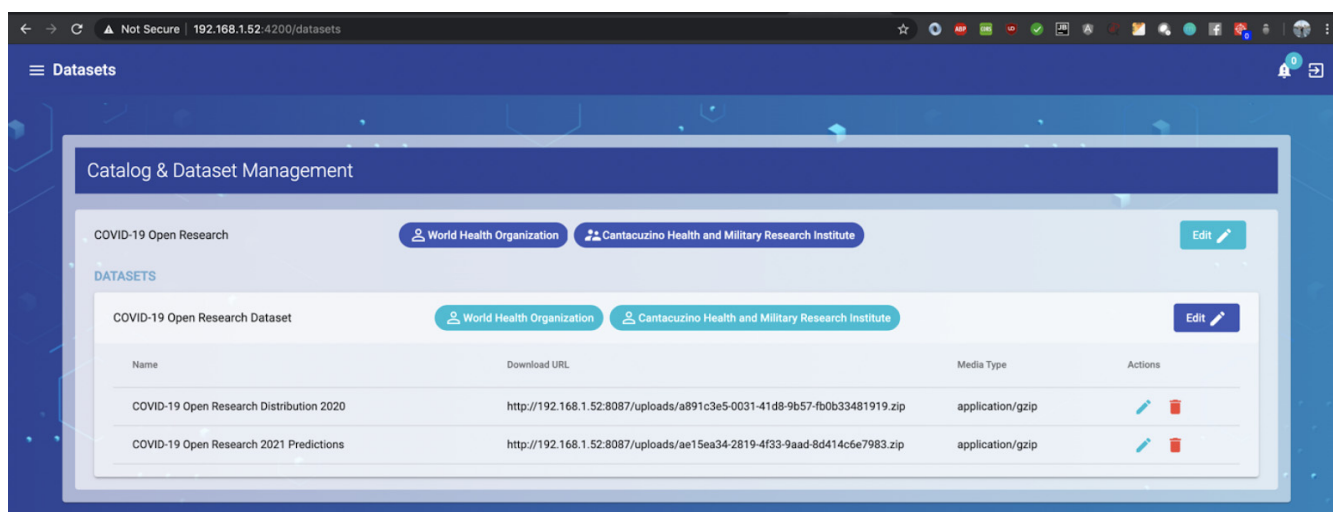


Figure 8. FAIR Metadata Management View (First Architectural Layer).

The edit view for a dataset is visible in Figure 9. As it can be observed, information about title, version, language, distribution license and categories of the dataset can be changed. Categories are added by querying either Wikidata [45] or DBPedia [44] for concepts that match with the highest accuracy the user provided input. Access rights can also be edited, providing the business accessing the functionality has control access. The metadata update process for catalogs and distributions is similar, although some properties in the semantic modelling may be different, as we can notice in Tables 1–3. These functionalities are encapsulated in the “Dataset Management” service, which runs the dataset, distribution, catalog and access policy smart contract to create new transactions on the corresponding assets from the following architectural level in Figure 4. On transaction approval from all the business nodes in the permissioned network, changes are committed on the Triplestore entries, using SPARQL queries.

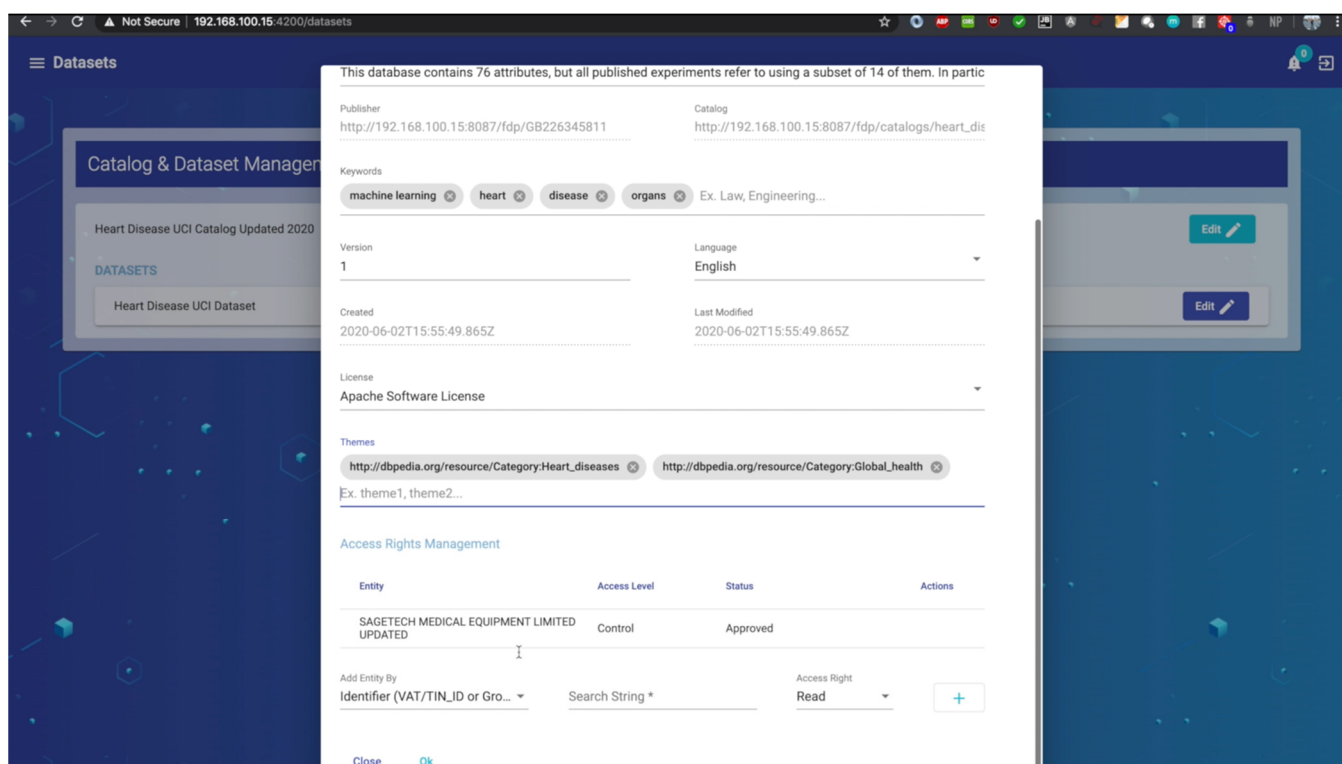
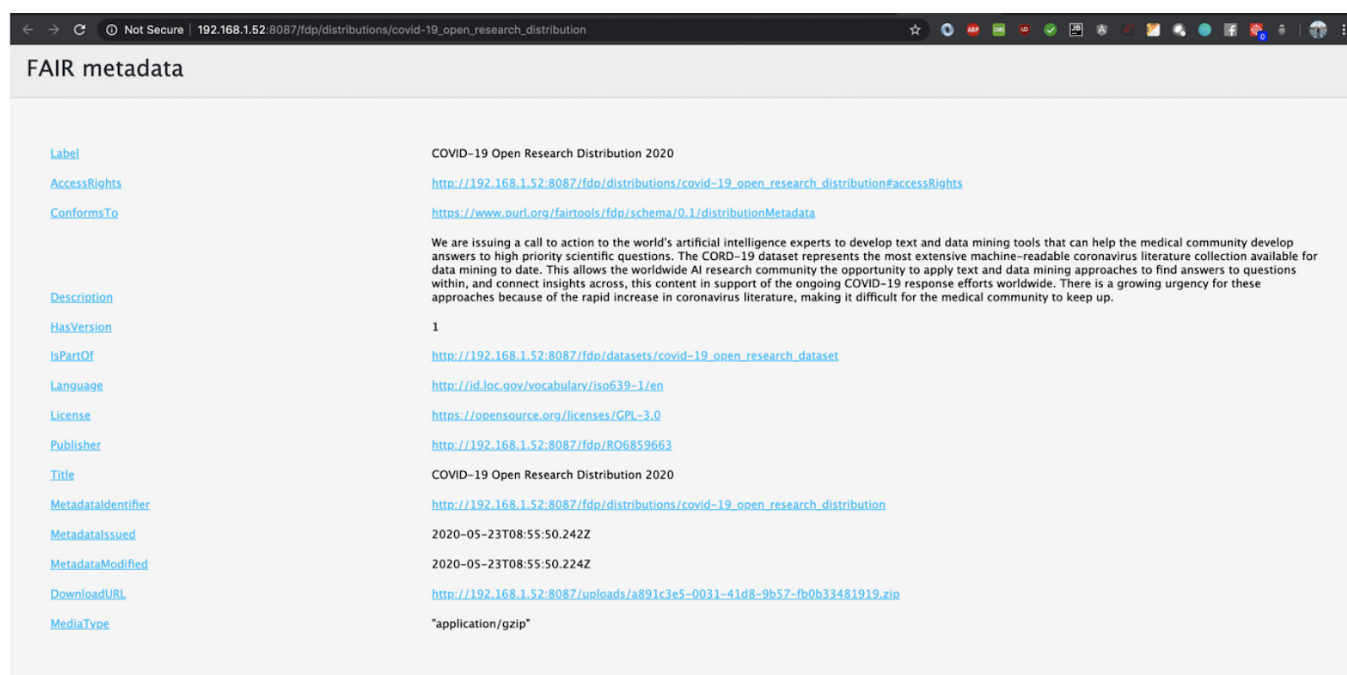


Figure 9. Dataset Edit View (First Architectural Layer).

6. P2P FAIR Data Points

In the FAIR specification, a data point is an API which returns the metadata associated with an URI representing a resource structured according to the proposed semantic model. In the context of DataShareFair, it was required to offer such an endpoint in order to make the metadata belonging to the catalogs, datasets, distributions, authorizations, access policies and business accounts available through an open protocol. This feature should facilitate the automatic processing and retrieval of FAIR metadata, often needed in B2B scenarios. In our solution, the FAIR data point specification is implemented in a distributed manner, since each data publisher offers its data through its own hosted endpoint in the network. As a result, the business owner has better control of usage over the shared data. In Figure 10, the HTML representation of the FAIR metadata of a distribution instance is shown. Its properties are semantic URIs linked to relevant concepts. Some of the values are linked to other resources belonging to the publisher, such as the dataset identifier, the download URL and the access rights policy. The FAIR data point accepts GET requests on valid routes, such as URI for existing catalogs, datasets and distributions. A SPARQL describe query over the received URI is run on the metadata collections stored in the Triplestore.



Label	COVID-19 Open Research Distribution 2020
AccessRights	http://192.168.1.52:8087/fdp/distributions/covid-19_open_research_distribution#accessRights
ConformsTo	https://www.oururl.org/fairtools/fdp/schema/0.1/distributionMetadata
Description	We are issuing a call to action to the world's artificial intelligence experts to develop text and data mining tools that can help the medical community develop answers to high priority scientific questions. The CORD-19 dataset represents the most extensive machine-readable coronavirus literature collection available for data mining to date. This allows the worldwide AI research community the opportunity to apply text and data mining approaches to find answers to questions within, and connect insights across, this content in support of the ongoing COVID-19 response efforts worldwide. There is a growing urgency for these approaches because of the rapid increase in coronavirus literature, making it difficult for the medical community to keep up.
HasVersion	1
IsPartOf	http://192.168.1.52:8087/fdp/datasets/covid-19_open_research_dataset
Language	http://id.loc.gov/vocabulary/iso639-1/en
License	https://opensource.org/licenses/GPL-3.0
Publisher	http://192.168.1.52:8087/fdp/RO6859663
Title	COVID-19 Open Research Distribution 2020
MetadataIdentifier	http://192.168.1.52:8087/fdp/distributions/covid-19_open_research_distribution
MetadataIssued	2020-05-23T08:55:50.242Z
MetadataModified	2020-05-23T08:55:50.224Z
DownloadURL	http://192.168.1.52:8087/uploads/a891c3e5-0031-41d8-9b57-fb0b33481919.zip
MediaType	"application/gzip"

Figure 10. FAIR Data Point Distribution Metadata.

4. Impact of DataShareFair in B2B Data Exchange Use Cases

A study on DataShareFair's impact over the current B2B sharing processes is presented relatively to the main decisional factors a company considers when engaging in such activities. The main impediments encountered by the European companies [1] are approached in the context of the proposed platform. Our main purpose is to show that the usage of DataShareFair in B2B sharing processes significantly reduces all costs. The direct contribution of the DataShareFair platform in the B2B data exchange processes is presented in Table 7.

Table 7. B2B Data Exchange Challenges in DataShareFair.

Category of the Decisional Factor	Decisional Factor in B2B Data Exchange Fulfilment	DataShareFair Approach on B2B Challenges	Overall Contribution
Legal and Economical	Liability avoidance through clear license agreements	When creating any type of resource such as a catalog, dataset and distribution meant to structure data through metadata entries, a license agreement must be selected from a predefined list; this ensures any business is aware of the legal scope of usage when crawling the metadata; data providers are protected from any wrongful reuse of their data	+
Technical	Operational barriers caused by interoperability and standardization issues	DataShareFair proposes structured metadata to be extracted from the shared data; metadata is modelled hierarchically according to the FAIR data principles and proposed ontologies; this helps businesses create suitable automated data discovery and reuse mechanisms	+
Technical	Poor or insufficient quality of the data	Extracted metadata is well-structured, split in granular fields, each serving their own purpose; by addressing the standardization issue with a strict data model, high data quality standards are also ensured	+
Technical	Infrastructure costs regarding storage, security and curation of available data	Security of the data is covered by the encrypted data storage in ZIP archives, which are decrypted only for the authorized businesses; the delete option is also available for the distribution files referenced in a dataset; Data can be cleared in order to reduce storage costs, but extracted metadata is kept for historical traceback;	+
Technical and Legal	Ownership rights assertion and definition of the legal extent to which data can be used	Businesses are able to acknowledge the existence of deleted data and they can contact the provider (through the business information provided in metadata) for subsequent access if needed Claiming the rights of ownership over uniquely created data is possible through automatic extraction of the publishing company's identity; the legal extent of data usage is specified by the owner throughout the attached agreement license, compulsory at the creation time of any dataset, catalog and distribution	+
Technical	Monitoring capabilities to control the extent of data usage	DataShareFair offers an integrated monitoring mechanism designed as an important feature of the PrivateSky platform; Each new transaction performed by an entity over data/metadata in the proposed marketplace is added to the blockchain log entries, where the business can obtain an overview over the means of usage for the provided data;	+
Economical	Costs associated to the skills development among employees within the company to analyze the available data	The proposed solution does not manage to fully cut the costs associated with training the personnel involved in the B2B data exchange process; It manages to slightly reduce the training time through a user-friendly designed tool which requires less training than going manually through unclassified TB of data;	—

Table 7. Cont.

Category of the Decisional Factor	Decisional Factor in B2B Data Exchange Fulfilment	DataShareFair Approach on B2B Challenges	Overall Contribution
Technical	Guidance, methodologies and systemic approaches for data sharing	<p>The proposed B2B data exchange solution establishes a well-defined procedure for both data providers and data users;</p> <p>Both parties need to be authenticated with real fiscal information (VAT ID and digital certificate), data is hierarchically organized based on the provided metadata and stored as encrypted ZIP archives in order to enhance security;</p> <p>data can be filtered and found based on the metadata information; access can be gained either through access requests or through direct authorization provided by the owner;</p> <p>Data can be managed only by the owners and people with minimum write access level;</p> <p>The trust component between two or many business entities is facilitated in DataShareFair through the access rights semantic model, applied over the shared data and metadata;</p> <p>A data provider can choose from different access levels when allowing another business to work with their data;</p>	+
Technical and Legal	Establishing trust with partners when sharing data	<p>Businesses can receive group access to resources if a joint legal agreement was put in place;</p> <p>The access policies management allows a data owner to also remove existing authorizations in case conditions of the legally binding agreement are broken;</p>	+
Technical and Legal	Ensuring compliance with regulations such as General Data Protection Regulation (GDPR)	<p>DataShareFair manages to ensure compliance with the GDPR legislation by offering sensitive data storage in encrypted blocks over the blockchain;</p> <p>When a business account is deleted, all private information such as email and phone number are erased from the system; only the unique FAIR identifier of the company is kept to ensure the consistency of the metadata semantic model;</p> <p>Distribution data provided by a business is GDPR compliant at the FAIRification step, prior to sharing;</p>	+
Technical and Economical	Addressing concerns that data sharing could result in a loss of business competitiveness or exposure of trade secrets	<p>A step forward in this direction brought by DataShareFair is encapsulated in the PrivateSky underlying technology; PrivateSky offers a mechanism to detect faulty nodes in the network through the OBFT consensus algorithm used for transaction validation;</p> <p>When an illegitimate node is detected, the data flow through the network is interrupted to avoid delivering it to unauthorized entities; the risk to expose important information to untrusted entities is significantly reduced;</p>	+
Technical	Denial or unforeseen termination of access to the datasets by the data supplier	<p>Businesses are not guaranteed unlimited access to existing distributions because of infrastructure costs on the data providers side; nonetheless, datasets, catalogs and distributions metadata cannot be deleted or hidden from marketplace by the data owners; this allows the data user to retrieve and mine relevant metadata and request access to data from the owner directly through the DataShareFair platform;</p>	+

Using DataShareFair in the B2B cross-sector data exchange process addresses most of the concerns businesses have shown over the years when met with the opportunity to engage in data sharing. The most important characteristics of our platform lie in its data security and privacy mechanisms, which are designed in respect to the proposed FAIR modelling of the data. As it was outlined in Figure 1, implementing the proposed B2B data sharing methodologies was possible by chaining a set of existing tools, such as Angular, PrivateSky, SPARQL and Apache Jena. The final result of our initiative is the DataShareFair, which offers data exchange mechanisms, while also ensuring data governance is not overlooked.

5. Conclusions

The current article, located at the interSection of B2B, Semantic Web, Blockchain and Privacy, brings both theoretical and practical contributions.

Following a comprehensive introduction in B2B data sharing, modelling and its impact in the business ecosystem in the first two sections, in Section 3 we propose a series of data governance methodologies to manage metadata structured according to our FAIR compliant ontologies. The resulting model can be traded by companies in the B2B data sharing platform which we described in Section 3, while in Section 4 we conclude over the impact of our proposal in B2B. The system's features rely on our own privacy and security centered blockchain mechanisms.

As for the future directions, DataShareFair should not only facilitate the B2B data sharing process, but also offer businesses opportunities to monetize their data. Given our blockchain based solution, this is possible by introducing wallet resources for each user into the current mechanisms for granting access to data. As a result, companies will be able to make their data available to others in exchange of being paid in various cryptocurrencies.

Another direction in our future work is to make the platform able to automatically translate legal formal contracts between partner businesses into authorizations and access rights over the data traded through our platform.

From a technical point of view, an aspect to improve is the FAIR extraction process, by introducing tokenization and NLP techniques to extract relevant intrinsic metadata. Using these techniques, we can build and continually improve a knowledge representation model based on our FAIR ontologies. This representation model can be used to compute appropriate solutions to specific questions a business has regarding the efficiency, scalability and profitability of its model.

Using DataShareFair as a platform for data sharing and reuse in the context of business to business processes would significantly improve their efficiency and reliability. The proposed solution manages to overcome technical impediments such as interoperability and discovery issues and also addresses several legal concerns, which result in a major reduction of the associated costs. As a result, more major players from significant industries would perceive B2B data sharing processes as the road towards innovative opportunities, rather than a costly investment with potential liabilities.

Author Contributions: Conceptualization, C.G.C. and L.A.; data curation, C.G.C. and L.A.; methodology, C.G.C. and L.A.; project administration, L.A.; resources, C.G.C. and L.A.; software, C.G.C.; supervision, L.A.; validation, L.A.; visualization, C.G.C.; writing—original draft, C.G.C. and L.A.; writing—review and editing, C.G.C. and L.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This research is partially supported by POC-A1-A1.2.3-G-2015 program, as part of the PrivateSky Project (P_40_371/13/01.09.2016) and by the Competitiveness Operational Programme Romania under Project Number SMIS 124759—RaaS-IS (Research as a Service Iasi).

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

References

1. Gartner. New B2B Buying Journey & Its Implication for Sales. Available online: <https://www.gartner.com/en/sales/insights/b2b-buying-journey> (accessed on 18 April 2021).
2. Gunnar, S. Business-to-business data sharing: A source for integration of supply chains. *Int. J. Prod. Econ.* **2002**, *75*, 135–146. [CrossRef]
3. Directorate-General for Communications Networks, Content and Technology. Study on Data Sharing between Companies in Europe (2018). Available online: <https://op.europa.eu/en/publication-detail/-/publication/8b8776ff-4834-11e8-be1d-01aa75ed71a1/language-en> (accessed on 20 April 2021).
4. Myler, L. Forbes. Available online: <https://www.forbes.com/sites/larrymyler/2017/09/11/data-sharing-can-be-a-catalyst-for-b2b-innovation/> (accessed on 13 September 2020).
5. Informatica. B2B Data Exchange—Streamline Multi-Enterprise Data Integration 2020. Available online: https://www.informatica.com/content/dam/informatica-com/en/collateral/brochure/b2b-data-exchange_brochure_6828.pdf (accessed on 23 April 2021).
6. Euro Banking Association. B2B Data Sharing: Digital Consent Management as a Driver for Data Opportunities. 2018. Available online: https://eba-cms-prod.azurewebsites.net/media/azure/production/1815/eba_2018_obwg_b2b_data_sharing.pdf (accessed on 20 April 2020).
7. Gartner. Magic Quadrant for Analytics and Business Intelligence Platforms. Available online: <https://www.gartner.com/en/documents/3980852/magic-quadrant-for-analytics-and-business-intelligence-p> (accessed on 11 February 2020).
8. Datapace. 2021. Available online: <https://datapace.io> (accessed on 18 April 2021).
9. Epimorphics. Epimorphics. 2021. Available online: <https://www.epimorphics.com/services/> (accessed on 18 April 2021).
10. iGrant.io. iGrant.io. 2021. Available online: <https://igrant.io/> (accessed on 18 April 2021).
11. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
12. Fair, R.T. FAIR Principles. 2020. Available online: [FAIRPrinciples](https://www.fairprinciples.org/) (accessed on 16 May 2020).
13. Association of European Research Libraries (A. o. E. R. Libraries). Implementing FAIR Data Principles: The Role of Libraries. 2017. Available online: <https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf> (accessed on 20 April 2020).
14. Commission, E. Guidelines on FAIR Data Management in Horizon 2020. Available online: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (accessed on 22 April 2020).
15. Gartner. Data Governance. 2021. Available online: <https://www.gartner.com/en/information-technology/glossary/data-governance> (accessed on 18 April 2021).
16. Talend. What is Data Governance. 2021. Available online: <https://www.talend.com/resources/what-is-data-governance/> (accessed on 19 March 2021).
17. European Data Portal. 2020. Available online: <https://www.europeandataportal.eu/en/training/what-open-data> (accessed on 22 April 2020).
18. European Legislation on Open Data and the Re-use of Public. 2020. Available online: <https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information> (accessed on 15 April 2020).
19. Link, G.J.; Lombard, K.; Conboy, K.; Feldman, M.; Feller, J.; George, J.; Germonprez, M.; Goggins, S.; Jeske, D.; Kiely, G.; et al. Contemporary Issues of Open Data in Information Systems Research: Considerations and Recommendations. *Commun. Assoc. Inf. Syst.* **2017**, *41*. [CrossRef]
20. Mohan, S. Building a Comprehensive Data Governance Program. Available online: <https://www.gartner.com/en/documents/3956689/building-a-comprehensive-data-governance-program> (accessed on 27 August 2019).
21. Calancea, C.G.; Alboaie, L.; Panu, A.; Swarm, A. ESB Based Architecture for an European Healthcare Insurance System in Compliance with GDPR. In International Conference on Parallel and Distributed Computing: Applications and Technologies; PDCAT 2018. In *Communications in Computer and Information Science*; Springer: Singapore, 2018; p. 931. [CrossRef]
22. Cowan, D.; Alencar, P.; McGarry, F. Perspectives on Open Data: Issues and Opportunities. In Proceedings of the 2014 IEEE International Conference on Software Science, Technology and Engineering, Ramat Gan, Israel, 11–12 June 2014. [CrossRef]
23. Fair, R.T. FAIR Data Point Specification. 2020. Available online: <https://github.com/FAIRDataTeam/FAIRDataPoint-Spec> (accessed on 18 May 2020).
24. Fair, R.T. FAIR Data Point Metadata Specification. 2020. Available online: <https://github.com/FAIRDataTeam/FAIRDataPoint-Spec/blob/master/spec.md> (accessed on 18 May 2020).
25. W3C. “WebAccessControl”. 2016. Available online: <https://www.w3.org/wiki/WebAccessControl> (accessed on 19 March 2021).
26. PrivateSky Project. 2020. Available online: <https://profs.info.uaic.ro/~jads/PrivateSky/> (accessed on 18 April 2021).

27. Alboaie, S.; Ursache, N.C.; Alboaie, L. Self-Sovereign Applications: Return control of data back to people. In Proceedings of the 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Verona, Italy, 16–18 September 2020.
28. Alboaie, S.; Alboaie, L.; Zeev, P.; Adrian, I. Secret Smart Contracts in Hierarchical Blockchains. In Proceedings of the 28th International Conference on Information Systems Development (ISD2019), Toulon, France, 28–30 August 2019.
29. Angular. 2020. Available online: <https://angular.io> (accessed on 15 May 2020).
30. Apache Jena. 2020. Available online: <https://jena.apache.org> (accessed on 16 May 2020).
31. Apache Jena Fuseki. 2020. Available online: <https://jena.apache.org/documentation/fuseki2/index.html> (accessed on 16 May 2020).
32. Docker. 2020. Available online: <https://www.docker.com/> (accessed on 20 May 2020).
33. Alboaie, L.; Alboaie, S.; Panu, A. Swarm Communication—A Messaging Pattern Proposal for Dynamic Scalability in Cloud. In Proceedings of the 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing, Zhangjiajie, China, 13–15 November 2013.
34. Alboaie, L. Towards a Smart Society through Personal Assistants Employing Executable Choreographies. In Proceedings of the Information Systems Development: Advances in Methods, Tools and Management (ISD2017 Proceedings), Larnaca, Cyprus, 6–8 September 2017.
35. Voshmgir, S. *Token Economy: How Blockchains and Smart Contracts Revolutionize the Economy*; BlockchainHub: Berlin, Germany, 27 June 2019; ISBN/EAN: 9783982103822.
36. PrivateSky. PrivateSky EDFs Explained. 2020. Available online: <https://privatesky.xyz/?API/edfs/overview> (accessed on 18 April 2021).
37. PrivateSky. What is Swarm Communication? 2020. Available online: <https://privatesky.xyz/?Overview/swarms-explained> (accessed on 18 April 2021).
38. PrivateSky. PrivateSky Secret Smart Contracts. 2020. Available online: <https://privatesky.xyz/?Overview/Blockchain/secret-smart-contracts> (accessed on 18 April 2021).
39. PrivateSky. PrivateSky Architecture. 2020. Available online: <https://privatesky.xyz/?Overview/architecture> (accessed on 18 April 2021).
40. PrivateSky. PrivateSky Interactions. 2020. Available online: <https://privatesky.xyz/?API/interactions> (accessed on 18 April 2021).
41. Calancea, C.; Miluț, C.; Alboaie, L.; Iftene, A. iAssistMe—Adaptable Assistant for Persons with Eye Disabilities. In Proceedings of the Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019, Budapest, Hungary, 4–6 September 2019.
42. DataShareFair Source Code. 2020. Available online: <https://bitbucket.org/meoweh/datasharefair/src/master/> (accessed on 15 July 2020).
43. VAT. Identification Numbers. 2020. Available online: https://ec.europa.eu/taxation_customs/business/vat/eu-vat-rules-topic/vat-identification-numbers_en (accessed on 10 April 2020).
44. DBpedia. 2020. Available online: <https://wiki.dbpedia.org> (accessed on 15 July 2020).
45. Wikidata. 2020. Available online: https://www.wikidata.org/wiki/Wikidata:Main_Page (accessed on 15 July 2020).