Open data infrastructures

Charalabidis, Yannis; Alexopoulos, Charalampos; Janssen, Marijn; Lampoltshammer, Thomas; Ferro, Enrico

DOI
10.1007/978-3-319-90850-2_6

Publication date
2018

Document Version
Final published version

Published in
Public Administration and Information Technology

Citation (APA)
Charalabidis, Y., Alexopoulos, C., Janssen, M., Lampoltshammer, T., & Ferro, E. (2018). Open data infrastructures. In *Public Administration and Information Technology* (pp. 95-113). (Public Administration and Information Technology; Vol. 28). Springer. https://doi.org/10.1007/978-3-319-90850-2_6

Chapter 6
Open Data Infrastructures

"User-centricity, transparency, and trust are the key elements towards a sustainable open data infrastructure."

6.1 Forming Open Data Infrastructure

Data represents a key asset in virtually any aspect of society and economy. Open Data in particular represents a source of immense value, as social capital (Lampoltshammer & Scholz, 2017) as well as an asset for business cases. Governments and their public administrations are generating and collecting during their service a plethora of different kinds of data, as well as an enormous amount in terms of volume. To tab into the potential this data holds in terms of stimulating economy, as well as the development and enhancement of governmental service for the benefit of the public (see Fig. 6.1), a sophisticate Open Data Infrastructure is required.

The Open Data Institute (ODI) sees data infrastructure as tangible and important as classical infrastructure, such as electricity or road networks. Data infrastructure have the main goal to keep the society informed and therefore contributes directly towards an increased accessibility and governance regarding data. Data within the infrastructure is quite heterogeneous, comprising not only governmental data, but also data from the business sector as well as data from non-profit organizations. The increased transparency in consequence can lead to not only business value, but also to environmental gains as well as towards societal benefits. In general, the ODI describes three different kinds of data infrastructure (Broad, Tennison, Starks, & Scott, 2015):

- **Local Data Infrastructure:** this kind of infrastructure contributes to an improved information state of citizens, communities, as well as decisions-makes on a governmental level
- **National Data Infrastructure:** this kind of infrastructure aims at strengthening the inherent resilience of a country in economic, social, and environmental areas.

Y. Charalabidis et al., *The World of Open Data*, Public Administration and Information Technology 28, https://doi.org/10.1007/978-3-319-90850-2_6

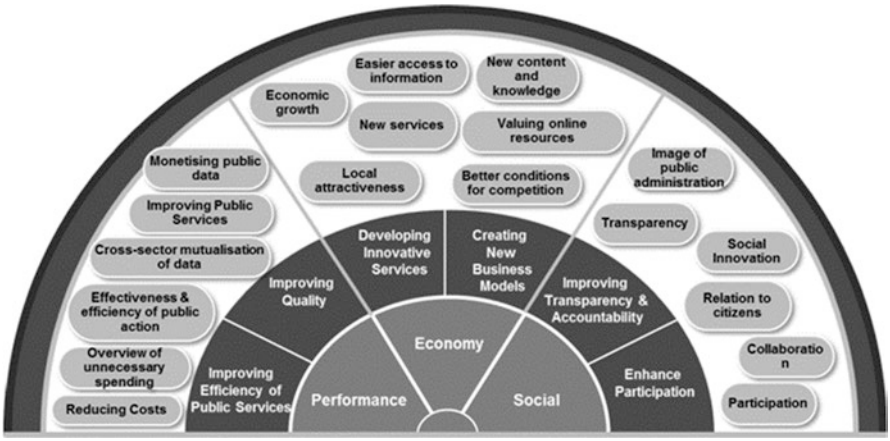

Fig. 6.1 Benefits of re-use of Open Government Data. (https://www.europeandataportal.eu/en/providing-data/goldbook/open-data-nutshell)

Besides the possibilities to build and provide services for citizens by companies and governments alike, the increased transparency boosts democracy as a whole.

- **Global Data Infrastructure:** this kind of infrastructure provides the means of tackling global issues such as getting insight to globally-acting entities such as multi-national organizations as well as a better understanding of progress regarding global policy-making.

With this important role of data infrastructure for individuals and society as a whole, there comes a great responsibility and requirements towards organizational and technological, as well as ethnical capabilities of organizations that provide these kinds of data infrastructure (Broad et al., 2015):

- **Long-term sustainability:** this kind of infrastructure contributes to an improved information state of citizens, communities, as well as decisions-makes on a governmental level,
- **Perceived authority:** citizens should hold a basic trust towards the maintainer of the data infrastructure, including its data,
- **Transparency:** the infrastructure should be transparent in a way that all processes regarding management and operations on the data themselves are well-documented and comprehensible, as well as replicable. Furthermore, the infrastructure should feature mechanisms, which allow for requests regarding an entities own data, what they were used for, who accessed them, etc.,
- **Openness:** the envisioned infrastructure should treat requests and users equally in terms of response, the right of information, as well as access to its inherent services and data, while at the same time protect the rights of individuals as required by law,
- **Commitment to the validity of data:** this attribution becomes most important in cases, where the infrastructure representing a de facto monopoly regarding

data storage and access to information. It should therefore be dedicated towards long-term sustainability of the data and information stored, paired with high aims regarding the provision of high-quality data, the use of standard formats, as well as its social responsibility towards the citizens,

- **Agility:** the infrastructure has to be able to not only handle the rapidly developing and demanding changes of technology and societal expectations, but also provide up-to-date data and information upon the request of external users.

All the before-mentioned criteria can be translated into a set of functional requirements, which should be fulfilled by an open data infrastructure. In the following section, we will have a closer look at these functional requirements, together with ongoing research activities regarding methodologies and tools to realize these requirements.

6.2 Functional Requirements of an Open Data Infrastructure

A sustainable open data infrastructure should reflect the needs and requirements of all involved stakeholders that are providing data to or using data from the data infrastructure. Zuiderwijk (2015b) conducted a research work towards the design of such an infrastructure to enhance the coordination of open data use. In particular, her study focused on the influential factor of OGD use, the functional requirements of an infrastructure for OGD, its functional elements, a concrete realization of such an infrastructure, and finally its overall effects. Table 6.1 provides an overview of the derived functional requirements of an open data infrastructure.

The requirements can be grouped within five main categories, namely, (i) searching and finding data, (ii) analysis of data, (iii) data visualization, (iv) interaction on this data, and (v) quality analysis of the data. In the following, we will have a look at current research works in these five respective categories.

Table 6.1 Functional requirements of an open data infrastructure

Category	Functional requirement
Searching and finding data	1. The OGD infrastructure should be a one-stop shop for datasets and metadata from a variety of other OGD infrastructures.
	2. The OGD infrastructure should allow OGD users to integrate and refer to datasets from various other OGD sources.
	3. Use controlled vocabularies to describe OGD.
	4. Use interoperable standards to describe OGD.
	5. The OGD infrastructure should support data search through keywords, data category browsing and data querying.
	6. The OGD infrastructure should support OGD use by the ability to search for data and metadata in multiple languages.
	7. The OGD infrastructure should facilitate filtering, sorting, structuring and ordering relevant search results.

(continued)

Table 6.1 (continued)

Category	Functional requirement
Analysis of data	8. The OGD infrastructure should provide data which describe the dataset.
	9. The OGD infrastructure should provide data about the context in which the dataset has been created.
	10. It should be clear for which purpose the data have been collected.
	11. It should provide examples of the context in which the data might be used.
	12. Domain knowledge about how to interpret and use the data should be provided.
	13. The OGD infrastructure should allow for the publication of datasets in different formats.
	14. The OGD infrastructure should offer tools that make it possible to analyses OGD.
	15. The OGD infrastructure should provide insight in the conditions for reusing the data.
Visualization of data	16. The OGD infrastructure should provide and integrate visualization tools.
	17. The OGD infrastructure should allow for visualizing data on maps.
Interaction on data	18. The OGD infrastructure should support interaction between OGD providers, policy makers and OGD users in OGD use processes.
	19. The OGD infrastructure should allow for conversations and discussions about released governmental data.
	20. The OGD infrastructure should allow for viewing who used a dataset and in which way.
	21. The OGD infrastructure should provide tools for interactive communications between OGD providers, policy makers, and OGD users (e.g. data request mechanisms and social media).
	22. The OGD infrastructure should provide tools for interactive communications between OGD users (e.g. discussion forums and social media).
	23. The OGD infrastructure should provide tools to keep track of amended datasets so that users know how datasets have been changed.
Quality analysis on data	24. The OGD infrastructure should provide insight in quality dimensions of OGD.
	25. It should be possible for OGD users, OGD providers and policy makers to discuss the quality of a dataset.
	26. The OGD infrastructure should provide information on the context in which a person reused a particular dataset.
	27. The OGD infrastructure should provide quality dimensions of datasets that are comparable with other datasets and with different versions of the same dataset.
	28. It should be possible to compare the quality of datasets over different data sources, over time and over data reuse on the data infrastructure.

Adapted from Zuiderwijk (2015a)

6.2.1 Searching and Finding Data

Sugimoto, Li, Nagamori, and Greenberg (2017) focused in their work on the topic of data archiving, especially metadata longevity. They provided suggestions and a proposed approach toward provenance of metadata registry in the area of risk management. In their work, the authors point out the challenges that arise from handling the context of the preserved metadata as well. This is a non-trivial problem, as the definition of concepts, which would be used to describe the context within a Linked Data environment, are prone to changes of time. Song (2017) proposed a method of linking data in the field of digital humanities across languages. This is achieved via use of metadata, yet without approaching the issue from the classical angle of translation. Instead, word embeddings are employed to then calculate a similarity metric based on the actual word vectors. The approach was successfully tested on a use-case involving Japanese and English. While there exists a plethora of shared vocabularies and ontologies, the actual engineering task of using them in a given context of a certain domain is challenging. Thus, precision regarding the description of concepts within an ontology is key. Out of this reason, Dutta, Toulet, Emonet, and Jonquet (2017) came up with a revised version of *Metadata vocabulary for Ontology Description and publication,* short MOD 1.2. This new version significantly increased the potential level of expressiveness of attribute-based ontology description, along with the possibility to semantic annotations via an OWL vocabulary to allow for the ontologies to be made available as Linked Data. When it comes to the task of creating Linked Data, e.g. in form of RDF, flexible and extensible tool are needed. To enhance current efforts in this research direction, Knap et al. (2018) introduced the UnifiedViews toolkit, an ELT framework that can handle a variety of associated processing tasks. Besides its capabilities of standard (pre-)processing tasks, custom modules can also be developed and integrated into the RDF creation workflow.

6.2.2 Analysis and Visualisation of Data

Kalampokis, Tambouris, and Tarabanis (2017) focused in their work on the combination of linked data approaches and open statistical data and the associated lifecycle. They created a toolkit named OpenCube, which allows for the associated actions specific to this data, covering its creation, expansion, and exploitation. Veith, Anjos, de Freitas, Lampoltshammer, and Geyer (2016) came up with a flexible, cloud-based solution for data processing and data fusion of heterogeneous sources, including open and closed data, based on a lambda architecture. By doing so, data of different temporal solutions and arrival speeds can be handled as well in various kinds of applications scenarios. The level of acceptance and adoption of open data strongly depends on the user experience delivered, while working and interacting with the data. An intuitive

representation is key in this circumstance yet is hard to achieve due to the high level of heterogeneity of Open Data. Thus, Ojha, Jovanovic, and Giunchiglia (2015) introduced a methodology, comprising a novel visualization approach, based on the concept of treating data as entities. This goes along with preferences of users to group and sort items by exactly such entities. Paired with a tailored UI, the authors could successful demonstrate the increased level of user experience, while browsing and searching through Open Data catalogues. Speaking of data heterogeneity, this becomes also an issue regarding the process of data integration. This heterogeneity is found via various formats (txt, csv, pdf), as well as the inherent schemata or not existing schemata. The work of Carvalho, Hitzelberger, Otjacques, Bouali, and Venturini (2015) discussed the pitfalls along the way of integrating this data, especially in the realm of Open Data. The authors show ways of dealing with the arising issues, stressing and demonstrating the pivotal role of information visualization to guide and support users in the integration task. A unique approach towards the visualization of "human-sensed data" is proposed by McLean (2017). She collected data concerning smells and aromas reported by citizens, while walking through the city. Combined with the geographic location of these reports, a visual olfactory map was derived, for communicating the results to the public. This interesting approach towards data visualization offers insights to citizen-collected data and lowers the barrier of comprehension of information.

6.2.3 Interaction on Data

Interaction and feedback loops regarding the data itself as well as the use of the associated services of the infrastructure from the public are imperative for sustainable platform. Thus, it is necessary to understand, how online communities can be incorporated into innovative co-creation processes to further evolve the existing offering of data and services. Konsti-Laakso (2017) for example focussed her research on two main aspects, namely how these online communities can help in drafting and executing innovation processes within the public sector and second, what kind of role social media platforms take in this process, including the produced results. Also, in the context of Smart Cities, technology and Open Data play an important role in the development and successful growth of the urban environment. However, the pure existence of data is not enough. Gagliardi et al. (2017) stress in their work the necessity of the data being used, feedback gathered, and also distributed and communicated. To enable this communication loop between citizens and government, the authors developed, based on a design science research methodology, an ICT-based tool name *UrbanSense*. This tool is envisioned to foster the innovation process of new public services, by enabling information flow even on a real-time level between citizens and public administration. When dealing with the cooperation of public administration and citizens, democratic processes represent important impact factors. Ruijer, Grimmelikhuijsen, and Meijer

(2017) argue that existing open data platforms are over-simplifying these processes and therefore have failed so far to hold up to their promises. To overcome this issue, they developed a *Democratic Activity Model of Open Data Use,* covering monitorial, deliberative and participatory use-cases, advocating a context-sensitive design approach towards data transformation and interaction. A special focus on the interaction with the Open Data community is put by the Austrian research project *ADEQUATe* (Höchtl & Lampoltshammer, 2016). Here, the project realised a community platform that provides enhanced versions of open datasets from the two main open data portals in Austria. The community is not only informed about the overall quality of data, but can also jointly work on the improved datasets, discuss related issues and changes, as well as provide further improved versions back to the community. For further details about ADEQUATe, please refer to Chap. 5.

6.2.4 Quality Analysis on Data

The overall quality of data is not only important in terms of reusability, but also towards credibility when it comes to open governmental data. Torchiano, Vetro, and Iuliano (2017) developed a basic set of metrics to assess open governmental contractual data, based on the ISO SQuaRE standard in a way that the fulfilment and potential problems within the data can be identified automatically. Stróżyna et al. (2017) developed a framework for identification of suitable open data based on quality and availability aspects to be combined with internal closed data to increase the overall values for an organization or company. The authors see restrictions, e.g. regarding automated crawling, as one of the most dominant hurdles, besides the general quality of the available data. Thus, the term Open Data should in their point of view being revisited, as it does not apply to various resources available on the Internet. Mihindukulasooriya, García-Castro, Priyatna, Ruckhaus, and Saturno (2017) also address the problem of data quality, yet from the specific viewpoint of Linked data. The developed a RESTful web service called Loupe API that provides profiling capabilities for Linked data based on user-specified requirements. These requirements can cover explicit details such as RDF classes or vocabulary, as well as implicit requirements such as cardinalities between entities and multi-lingual aspects. The results can of their API can either be inspected manually or via dedicated validation languages such as SPIN. Further information regarding data quality metrics and assessment can be found in Chap. 8.

Besides all functionalities of a platform or data infrastructure, it will not reside without the trust of the users regarding the process being correct, the data hosted being valid, as well as their individual rights being protected. Thus, the next section puts its focus on the important aspect of trust and how modern technologies can enable trust in open data infrastructures.

6.3 Building Trust in Governmental Data Infrastructures

Trust in the governmental domain can be visited from two perspectives. The first perspective relates to the trust of citizens towards the public administration. If citizens trust the processes they are involved in, less feedback and personal interaction is required, which can result in reduced overhead and thus in less cost and time. The other perspective is the one of the public administration where monitoring and validating actions, documents, and information provided by citizens take time and produce costs as well (van de Walle, 2017). So, in order to approach trust form the viewpoint of both parties, a common technology-based approach to be incorporated into the data infrastructure has to be found. As one solution towards this issue, we will discuss the concept and applicability of blockchain technology.

6.3.1 Transparency Through Blockchain Technology

The overall concept of blockchain is basically a kind of database, which is hosted over a network infrastructure (e.g., Internet) in a de-centralised and distributed way (Ølnes, 2016). In particular, a blockchain is not only storing but on the same time updating all transactions that it stores over all connected nodes within the P2P network. On this network, all nodes can make use of it to store their transactions, with every party receiving its own copy of the transaction. It is noteworthy that nodes do not have to be actual human users but can be – along the paradigm of IoT – also machines and software services. Signing-up to this distributed ledger is possible via the use of public key algorithms. The validation is performed by all nodes, to build a consensus about the correctness of the submitted transaction. If a transaction is declared valid, it is stored within a block, which in return is added to the blockchain. Thus, the last added block also states the trust of the network towards the correctness of the current chain. Every block is a set of transactions including associated timestamps, as well as the hash of the previous block within the chain (Ølnes, Ubacht, & Janssen, 2017). A simplified summary of the main steps within a blockchain transactions can be seen in Fig. 6.2. Blockchain technology was also successfully introduced as secure information management and provenance infrastructure throughout several countries with a strong e-government background (Ojo & Adebayo, 2017). To dive deeper into the context of blockchain technology in the public sector, the following section presents benefits and application scenarios in this very domain.

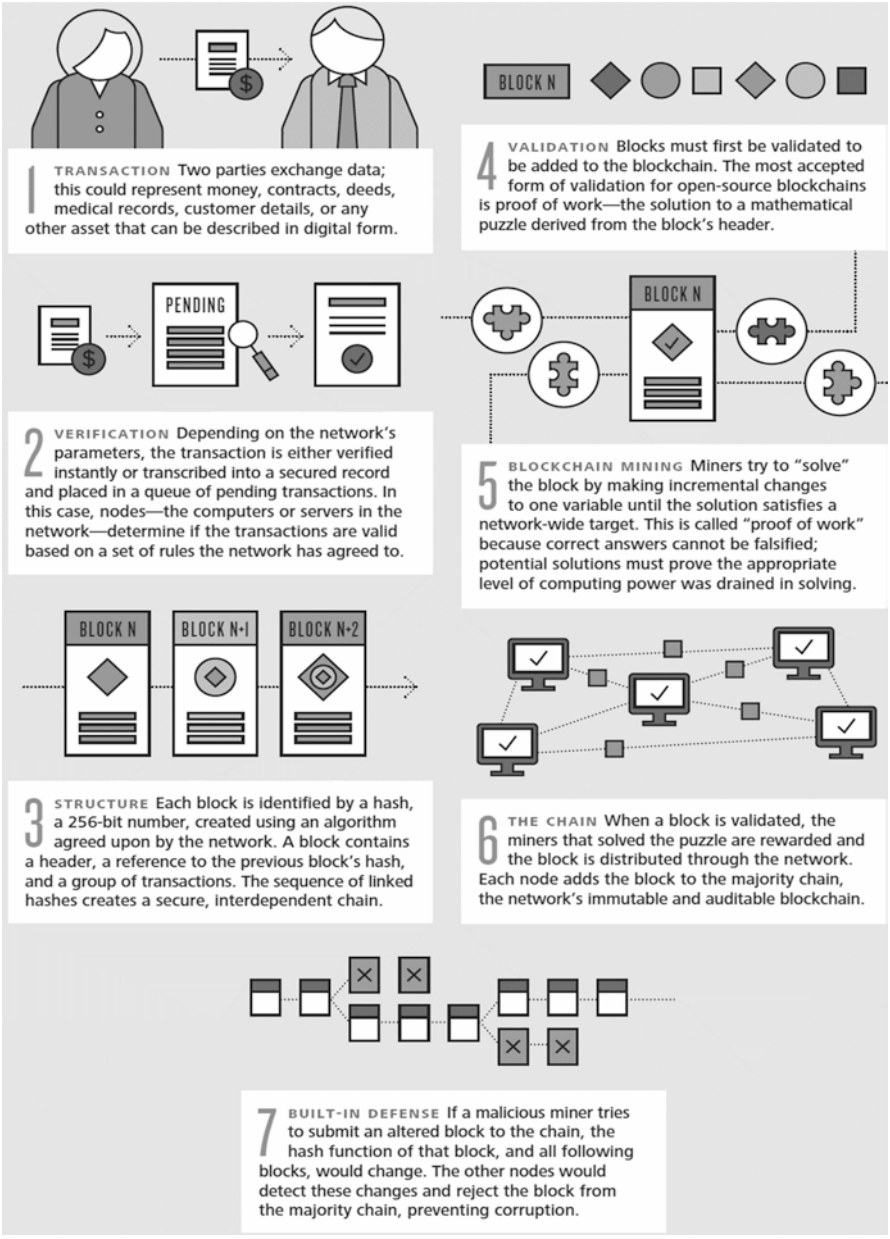

Fig. 6.2 The principles of a blockchain workflow. (Adapted from Piscini, Guastella, Rozman, and Nassim (2016))

6.3.2 *Benefits and Applications of Blockchain Technology in the Public Sector*

The benefits that can arise from blockchain technology are manifold and can range from strategic aspects, over organisational aspects, to economic aspects. Ølnes et al. (2017) provide a comprehensive overview over these aspects as can be found in Table 6.2.

The before-described features of blockchain technology demonstrate the great potential of its application in numerous scenarios. Considering this technology in the governmental sector, the following application use-cases can be identified (Fig. 6.3) (Welzel, Eckert, Kirstein, & Jacumeit, 2017):

- **E-Payment:** blockchain technology is best-known for its applicability in payment systems (e.g., bit coin). Therefore, it could also be used to make payments towards the government and vice versa. Examples here could be tax payments/

Table 6.2 Benefits and features of BC in governmental application

Category	Features	Description
Strategic	Transparency	Democratizing access to data. History of transactions remains visible and every node has complete overview of transactions.
	Avoiding fraud and manipulation	Hacks or unauthorized changes are difficult to made without being unnoticed, as information is stored in multiple ledgers that are distributed.
	Reducing corruption	Storage in distributed ledgers allows for preventing corruption. For example, by storing landownership in a BT and having clear rules for changing ownership which cannot be manipulated.
Organizational	Increased trust	Trust in in process by increased control due to immutable recordkeeping and by verification of the data by multiple nodes.
	Transparency and auditability	Being able to track transaction history and create an audit trail. Also, by having multiple ledger which can be accessed for consistency.
	Increase predictive capability	As history information can be traced back, this availability of the historic information increased the predictive capability.
	Increased control	Increased control by needing consensus to add transactions.
	Clear ownerships	Governance need clearly defined and how information can be changed.
Economical	Reduced costs	The costs of conducting and validating a transaction can be reduced as no human involved is needed.
	Increased resilience to spam and DDOS attacks	Higher levels of resilience and security reduces the costs of measure to prevent attacks

(continued)

Table 6.2 (continued)

Category	Features	Description
Informational	Data integrity and higher data quality	Information stored in a system corresponds to what is being represented in reality due to need for consensus voting when transacting and distributed nature. This result in higher data quality.
	Reducing human errors	Automatic transactions and controls reduces the making of errors by humans.
	Access to information	Information is stored at multiple place which can enhance the easy the access and speed of access.
	Privacy	User can be anonymous by providing encryption keys or access can be ensured to avoid others to view the information.
	Reliability	Data is stored at multiple places. Consensus mechanisms ensures that only information is changed when all relevant parties agrees.
Technological	Resilience	Resilient to malicious behaviour.
	Security	As data is stored in multiple databases using encryption manipulation is more difficult. Hacking them all at the same time is less likely.
	Persistency and irreversibility (immutable)	Once data has been written to a BC it is hard to change or delete it without noticing. Furthermore, the same data is stored in multiple ledgers.
	Reduced energy consumption	Energy consumption of the network is reduced by increased efficiency and transaction mechanisms.

Adopted from Ølnes et al. (2017)

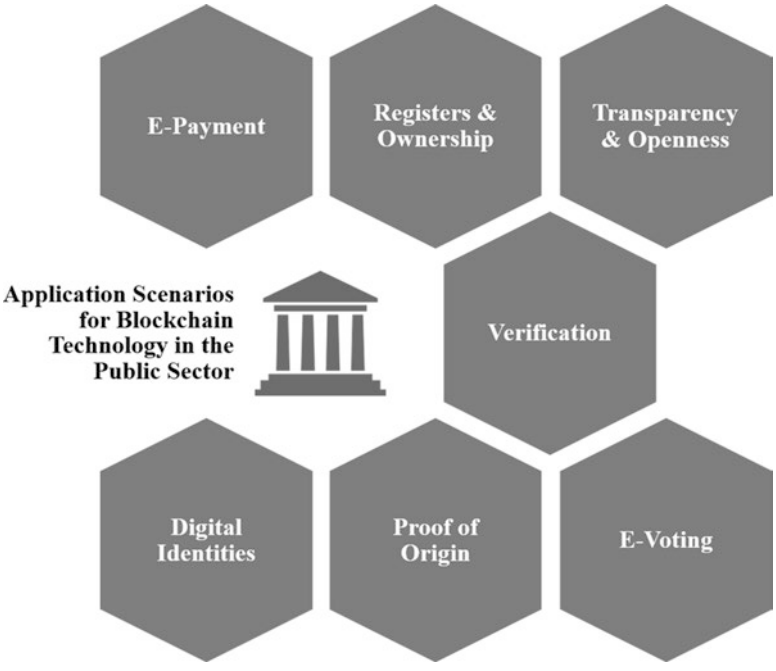

Fig. 6.3 Blockchain application scenarios. (Adapted from Welzel et al. (2017))

refunds, fees for certain services, as well as fines for violations. But not only monetary transfers between the citizens and the government, but also payments within the government as organization could be covered. These would include payment of salary, food stamps, parking tickets etc.

- **Registers and Ownership:** public registers, legal titles, as well as cadastres are common application examples for blockchain technology. The blockchain provides with its inherent transparency and immutability the means to prevent corruption, manipulation of existing entries, as well as a straightforward transfer of ownership. Furthermore, BCT can enable and enhance between governmental organizations on a national but also on an international level regarding the exchange of information, documents, and the verification of the existing of these documents.
- **Verification:** Verification of documents and data as well as their integrity are usually achieved via the use of digital signatures. This technology is established and currently used throughout different domains, including the governmental sector. Yet, they add an additional level of overhead to the process. First, there is the need for a central, trusted authority that issues the signatures and thus confirms the identity of the person acquiring the signature. Second, in order to be able to work with the signature, additional devices and/or software components are required, which add additional costs as well as might block certain application scenarios. BCT could help to reduce the burden of document verification and therefore increase the speed of the overall process.
- **Proof of Origin:** BCT can provide benefits in scenarios, where the traversal of a product through a process, e.g. a supply chain, has to be monitored in a way that every step can be verified. This can contribute to the fulfilment of legal compliance requirements. The public administration can also tap into this potential in cases, where it has the responsibility to govern over critical product/process flows, such as food chains or the trade with rare goods such as diamonds or art pieces.
- **Digital Identities:** the integrity of a digital representation of all ID relevant attributes can also be verified via a blockchain, by hashing all relevant attributes and storing the hash values within the chain. This concept could even be pushed further to use it as a kind of single-sign-on (SSO) system for organizations by including access rights to systems and services. The chain can then be used to check, if a person is allowed to access the particular service, system, or files. In addition, changes to the rights (withdrawal of rights or the addition of rights) can be seen via the history of changes within the blockchain.
- **Transparency and Openness:** today's society is demanding for transparency regarding the processes and actions taken by the government. Blockchain technology can help to provide this transparency and therefore contribute to the increase of overall trust of society towards its government and the elected representatives. A good example can be found in open data portals, which release open governmental data to the general public. By using BCT, the origin and integrity of this data can be verified, again improving trust towards the released information from the government, including accountability. Another example could be the budget of a government or parties, revealing all transactions and

spending, including donations and in consequence, making any lobbying activities and potential bias transparent.
- **E-Voting:** the matter of electronic voting is being discussed from various viewpoints, starting from e-voting being already implemented up to being completely anti-e-voting. Besides the principle "yes or no" discussion, BCT can be used for the voting process. Similar to the concept of bitcoin wallets, political candidates or parties could be equipped with a digital wallet and each citizen could vote with his or her specific single token towards the candidate or party. The candidate/party with the most tokens within their digital wallet win the election. While from a technical standpoint this is convenient, the approach also includes several caveats such as giving up to some degree anonymity of votes or could encourage tactical voting (as the number of votes are instantly visible) as well as potential bribing for securing the pivotal votes.

6.4 Real-World Examples of Open Data Infrastructures

6.4.1 Industrial Data Space

The German project Industrial Data Space (IDS) is one example of an open data infrastructure, with a particular focus on industrial applications. The IDS is based on the following core principles (Otto et al., 2016):

- **Data sovereignty**: the control over data within the IDS is never given up by the owner of the data. Thus, it is possible to link the data with licensing/terms and conditions that regulate operations with this data.
- **Secure data exchange**: a dedicated layer offers secure exchange of data between two or several entities, not only on a point-to-point bases, but also throughout complex supply chains.
- **Distributed architecture**: the IDS interconnects via its IDS connector all endpoints towards a distributed net of participants, without the necessity of a central authority or single-point-of-failure. The exact type of the architecture is set by the application scenario and is driven by economic aspects, specific to the market and domain at hand.
- **Data governance**: as described before, there is no central authority within the IDS. Therefore, participants of the IDS have to agree to a common rule set of how to work together, including duties and responsibilities. While this can be tricky to find common ground, at the same time, it provides the necessary flexibility to open the IDS for any application scenario and domain.
- **Network of platforms and services**: as the IDS is embracing the paradigm known as "Internet of Things" (IoT), the role of a Data Provider is not only limited to individuals or organizations, but can be also taken by devices, e.g., production machines, vehicles, etc. In additions, other Data Spaces/Markets can also interact with the IDS, and therefore with its entire ecosystem of stakeholders.

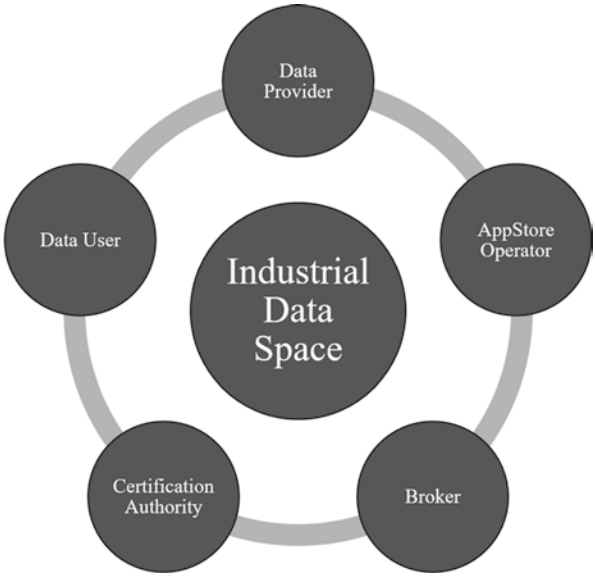

Fig. 6.4 Role concepts within the Industrial Data Space

- **Trust within the IDS**: without a common level of trust within the data space environment, participating actors will not engage with each other in terms of data exchange as well as service consumption. It is for this reason that participation is only possible by using the IDS connector, providing the required means of authentication and authorization.

While main goal of the IDS is the facilitation exchange between Data Providers and Data Users, other actors take important roles within this facilitation process (see Fig. 6.4). The actor environment within the IDS allows for a participant to enact several roles, including the possibility to rely on third parties for fulfilling tasks on their behalf. In the following, the distinct roles and their function within the environment of the IDS are explained (Otto et al., 2016).

The **Data Provider** holds the access to the sources from which data is offered towards the other participants of the IDS, while the data provider always keeps the control over the data. Furthermore, it offers descriptive information for the Broker to be able to properly register the data and offer it to interested stakeholders/actors throughout the IDS. The Data Providers is also responsible for the entire processing of the data within the IDS, including required transformations according to the inherent data model of the IDS, along with any applying terms and conditions in regard to the data itself. Finally, the Data Provider also orchestrates requests of data, in conjunction with handling the entire app and service ecosystem of the IDS. The role of the **Data Users** within the IDS is based on the consumption of data and services/apps, provided by other actors (Data Providers). This can either be a single source or multiple sources, including the required transformational as well as

mapping-based actions, which are required to achieved compatibility with the targeted data model.

The **Broker** functions as intermediary, bringing together the searching party (Data Users) with the providing party (Data Providers). Furthermore, the Broker acts as central register for data sources within the IDS. Thus, the Broker also handles services such as the provision of means for Data Providers to publish their data, as well as the provision of search and retrieval capabilities for the Data Users to browse the registered data sources. In consequence, the Broker also facilitates the creation of agreements and the associated provision of the data between the involved parties. The exchange of data is therefore supervised and recorded to ensure a secure and complete transaction. This also includes potential rollbacks in case that a transaction failed. As the Broker is a central role within the exchange of data, it can also be set up to offer supplementary services to all involved parties such as quality assessment of data, or additional analytical services. The **AppStore Operator** holds the central authority regarding 3rd-party software, developed by participants to be distributed within the digital business ecosystem of the IDS and its AppStore. Therefore, the AppStore Operator provides means of describing and registering software to be offered to customers, including the download of these services, as well as payment functionality and rating options for the offered software services. Finally, there is the **Certification Authority** which exists to ensure that all components on the IDS meet the jointly-define requirements of all participants. This includes activities such as the handling of the entire certification process, starting from the request up to the approval/denial of the certification, operation of the reporting system of testing parties, up to the issuing of actual certificates. To guarantee a consistent, fair, and comparable process, the Certification Authority maintains a criteria catalogue, which acts as basis for the certification process.

To demonstrate the feasibility of the concepts inherent to the IDS, the following use cases are developed and realized:

- **Truck and cargo management in inbound logistics:** supply chains often suffer from the fact that data is unnecessarily duplicated by involved companies, thus causing storage and synchronization issues between each particular stage of the chain. This results in higher costs due to increased processing and slower or even delayed delivery. Therefore, an increased level of transparency is required, enabling consistent monitoring throughout the entire supply chain and thus, improving transportation as well as quantitative and qualitative forecasts. A good example for the before-mentioned situation can be found in truck and cargo management. In order to guarantee an efficient and effective management process, it is crucial for all relevant data to be available once the truck arrives at its destination for follow-up tasks (e.g., check-in, job order planning). Yet, this data is not always available in a complete form, due to, e.g., different freight carriers employed by the shipping companies. The IDS will solve this issue by the introduction of suitable standards and a general simplification of the data exchange process (i.e., data regarding the order itself, data about the transportation such as GPS data, master data of suppliers).

- **Development of medical and pharmaceutical products:** for medical and clinical data being highly-sensitive due to its personal aspects, it is also highly-heterogenous, as it consists of data by individuals, institutions, and machines. Also, this kind of data is due to its sensitive character rarely aggregated within one single place. Thus, this fact also represents a hurdle within the process of developing new treatments, therapies, and medication. But availability alone is not enough, information about the context of the collection process, as well as the involved IT systems, and the overall quality of the data itself are imperative to generate a complete picture. To overcome these issues, IDS will provide means for aggregation of data, as well as the required transformations to enable analyses. This will not only strengthen ongoing studies, but also allow for hypotheses testing beyond existing scales and flexibility. Combined with the open and standardized interfaces of IDS, various systems can be interconnected to enhance processing, visualization, and exploration of data. Furthermore, anonymization services will provide the requirements defined by law to fully-comply with GDPR and associated laws and regulations to protect privacy.
- **Collaborative production facility management:** modern production environments require a high level of data completeness, e.g., regarding individual components, utilization of machines, material availability etc. Currently, costs regarding the collection, analysis, and distribution of this data are high, as often this data cannot be collected by the companies own capabilities and therefore requires 3rd party support. While developments regarding standardized BUS-based systems have improved over the years, interconnectivity and data exchange represent still challenging tasks, especially while the entire sector faces an intense and ongoing transformation due to the Internet of Things (IoT) paradigm. This becomes even more obvious when considering the task of transferring sensitive company data towards the company's own perimeter. The IDS can step in at this point to act as a pivotal point regarding the secure and standardized exchange of data between different parties, especially across organizational borders. Furthermore, IDS can provide companies acting as participants additional services, which can support the companies in performing analyses on their data, which they have not been able to do before. Finally, manufacturers could offer their data on the IDS as well, opening new business models as well as to establish grounds for new cooperation between participants.
- **End-to-end monitoring of goods during transportation:** in certain domain, transportation of highly-critical goods requires for special monitoring during the transportation process to avoid damage or destruction of the goods themselves. Examples can be found in form of electronics, medical supplies, or chemicals. These damages can occur due to high and/or rapid temperature changes, shock/vibration, light exposure etc. Potential countermeasure come with, e.g., sensors that can communicated changes in the environment the goods are currently traveling, or the status and condition of the goods themselves. The IDS enables secure and complete end-to-end monitoring of the

transportation, informing customers and suppliers alike in case something should be wrong with the goods. The IDS therefore covers an important aspect of future IoT applications.

6.4.2 Data Market Austria

As the volume of data in our today' society is growing by the minute, it is more than natural for it to be considered an important "raw material" throughout all industrial and business sectors alike. In consequence, an effective and efficient ecoservice for handling this data within the Austrian economy is an imperative factor for sustainability regarding business and society as a whole. Currently, there is no agenda regarding such an ecosystem for Austria and ongoing initiatives are still working towards a significant breakthrough. While platforms regarding, e.g., governmental open data and open data from business exist, they are not connected, and business use cases have no common platform as a central host around this data. Yet, even with the data available, it often lacks a common data quality standard and thus suffers from interoperability issues. Therefore, the Data Market Austria (DMA) is trying to overcome these issues by performing the following actions[1]:

- **Advancing Technology Foundations**: this roadmap foresees three distinct steps. In the first one, blockchain technologies is used to incorporate a decentralized way of security of data registration, computation, as well as provenance. The second step builds upon brokerage services, including the use of sophisticated recommendation algorithms for an improved match of users and data/service providers. The third step ensures the provision of all required timely computational capabilities for all operations on the market, including the fusion of different data sources.
- **Creating a Data Innovation Environment**: DMA strives for the inclusion of various stakeholder groups, starting from start-ups and SMEs, large enterprises, academia, up to public administration. This will build an interaction and innovative environment on a co-creation basis, which allows for a variety of business models, guaranteeing the flexibility to provide long-term sustainable solutions for all involved parties.
- **Cloud-based infrastructure**: the DMA will host its services in a cloud environment, thus providing a transparent and highly-scalable service infrastructure for all participants and their individual use cases, applications and business models.

The Data Market Austria envisions similar roles as the Industrial Data Space. An overview of the seven different roles can be seen in Fig. 6.5. One of the most significant difference of the DMA and IDS is that DMA is not mainly focusing on the industrial sector, but aims to bring together stakeholders of different domains, sizes of companies as well as public administrations and actors from academia.

[1] https://datamarket.at/ueber-dma/

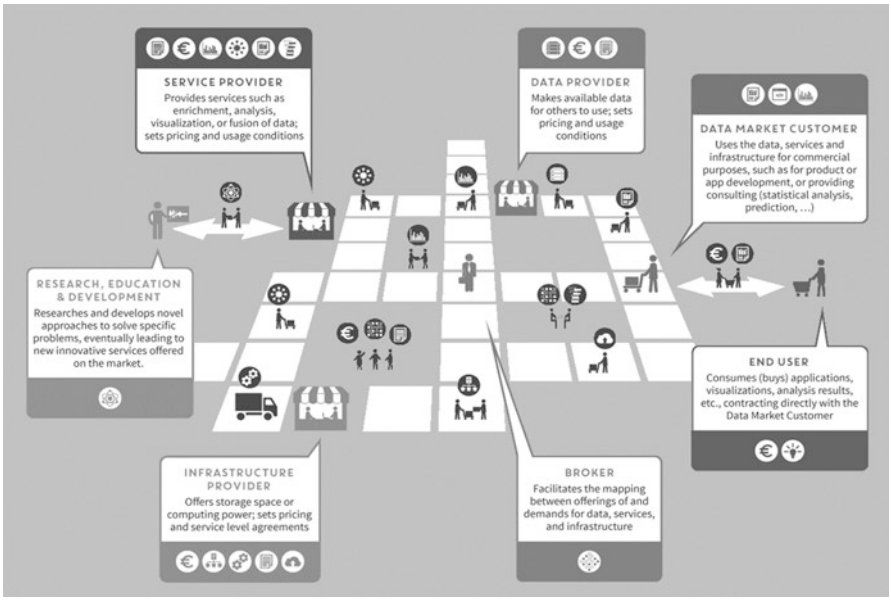

Fig. 6.5 Stakeholders within the Data Market Austria. (https://datamarket.at/en/)

To test the newly developed infrastructure, not only regarding technological requirements, but also in consideration of different business aspects of different domains, two main use case scenarios are covered within the DMA:

The first use-case is dedicated to the field of **Earth Observation**[2]. Here, data providers shift their access schemata more and more towards open access. An example in the European realm can be found with the Sentinel missions. The European Space Agency (ESA) and the European Commission are following a new set of policies, providing free access to this satellite data for any entity interested. Thus, earth observation becomes more and more accessible to the public as well as the industry sector for being used in business cases. Yet, single access to information sources is often only partly covering the requirements of Data Users, and linkage of data of different Data Providers is of great importance. For this reason, the DMA foresees Earth Observation data and services to be hosted on a cloud platform, allowing user to share infrastructure and data alike. This is in-line with the ESA's current activities regarding the process of distribution of Sentinel satellite data via a network of distributed data hubs. The DMA will demonstrate its capabilities in the EO domain, along with the linking of geodata on several application scenarios in the area of forestry.

The second use-case is placed in the **Mobility**[3] sector. Current data storage solutions are not suited for processing millions of data events transmitted, as they

[2] https://datamarket.at/earth-observation/

[3] https://datamarket.at/mobility/

can occur within, e.g., an IoT environment. Thus, a high level of scalability is imperative for future industrial but also other related business cases. Connected mobility solutions do present such as business case and application domain. Within this area, real-time prediction is considered one of the most time-consuming and computational demanding tasks. Thus, DMA will demonstrate its feasibility on two application examples in this field. The first example is dedicated to the task of **Taxi Fleet Management**. Here, public data and proprietary data will be used to optimize planning of taxi placement. Examples for data to be used are public transportation data such as arrival times of planes and trains, weather forecasts, local events, and mobile phone data of users that opted-in to make this data available. The second example comes in form of **Historical Traffic Flow Characteristics**. It is intended to derive patterns from historical data regarding traffic flow and mobility preferences of customers. This kind of data and predication could not only be of interest to taxi fleets but also towards city and urban planners to optimize traffic concepts as well as other related process towards improved traffic characteristics of the entire city.

6.5 Conclusion

In this chapter we have discussed the importance of open data infrastructures for a society, from both perspectives, the economic perspective as well as the governmental perspective. We have seen the high level of functional requirements that have to be fulfilled, in order to develop a sustainable infrastructure for open data and data in general. One of the most important aspects is present in the requirement of transparency and trust of the citizens towards the infrastructure as well as of the governmental organisations towards their potential users. State-of-the-Art technologies such as blockchains can help to provide the required level of transparency, while being open towards a variety of use-cases. We have discussed several use-cases in the domain of public administration and have seen that some of these could be realized already today, while for others it has still to be seen, if they can survive the scepticism of all involved parties as well as current legal obligations. While maybe in the public sector its advent is still not quite there, for sure it is becoming more and more present in the economic domain. In combination with open data, this has the potential for a huge variety of profitable business models. For more information on the value chain of open data and associated business models, please continue towards Chap. 7.