Organizational issues

How to open up government data?

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

DOI
10.1007/978-3-319-90850-2_4

Publication date
2018

Document Version
Final published version

Published in
Public Administration and Information Technology

Chapter 4
Organizational Issues: How to Open Up Government Data?

When publishing data, governmental organizations are often hindered by issues such as the lack of standard procedures, the threat of privacy violations when releasing data, the risk of accidentally releasing policy sensitive data, the risk of data misuse, and problems with data ownership.

4.1 Introduction

Governments create and collect enormous amounts of data, for instance concerning voting results, transport, energy, education, and employment. These datasets are often stored in an archive that is not accessible for others than the organization's employees. To attain benefits such as transparency, engagement, and innovation, many governmental organizations are now also providing public access to this data. However, in opening up their data, these organizations face many issues, including the lack of standard procedures, the threat of privacy violations when releasing data, accidentally releasing policy-sensitive data, the risk of data misuse, challenges regarding the ownership of data and required changes at different organizational layers. These issues often hinder the easy publication of government data.

In Chap. 2 we already discussed the open data lifecycle, including the steps that organizations take in opening data. This chapter discusses these steps and their related issues and potential effects more in depth. In this chapter we first discuss issues that governmental organizations face when opening up their data. We give an overview of all the issues, including the potential positive and negative effects, and then discuss each of them in detail, with a related example from the open government domain. Subsequently, we provide a use case that describes solutions to overcome some of the outlined issues. Thereafter, we describe best practices that function as guidelines for governmental organizations that want to open up their data. Such guidelines can be used by public organizations to improve their open data publishing processes. Ultimately, the implementation of the guidelines reduces barriers, stimulates the publication of government data, and contributes to attaining

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_4

the benefits of open data. Discussions with practitioners showed that the guidelines could improve the open data publication process.

4.2 Organizational Issues for Opening Up Government Data

Let us imagine that you are a civil servant working for a governmental organization, for instance, a ministry. As part of your daily tasks at the ministry, you have collected a number of datasets, and you consider opening the collected data. Which aspects do you need to consider? The main issues that public organizations may face when opening up their data are depicted in Table 4.1 (adapted from Janssen, Charalabidis, & Zuiderwijk, 2012; Susha, Zuiderwijk, Charalabidis, Parycek, & Janssen, 2015; Zuiderwijk, 2015a; Zuiderwijk & Janssen, 2015; Zuiderwijk et al., 2012b). We provide an example of each organizational issue and explain these issues further in the following sub sections.

Table 4.1 Organizational issues for opening up government data

Types of organizational issues	Organizational issues	Example
Data related issues	Potential privacy breaches	The Ministry of Justice collects data concerning crime victims and offenders. The data may be of interest to the public, yet it can only be opened up after it has been anonymized and/or aggregated.
	Data sensitivity and security	Data collected by the Ministry of Education may be sensitive, since it contrasts information provided by the responsible Minister of Education.
	Embargo period	A researcher working at a ministry first wants to publish an article and a report using the collected data. The data can only be opened after the article and report have been published.
	Data openness, lack of control over its use and lack of trust in the data user	A dataset concerning employment has been published online. After publication, the dataset is copied into various online repositories. Although this enhances openness, it is not clear to the data provider anymore at which places the data is available and how it is used. The data may be misused.
	Data quality	Some datasets are of high quality (i.e., they are complete, accurate, timely, and reliable), whereas for some datasets, the quality is low (e.g., the dataset is not complete) or it is unknown what the quality level is.
	Data documentation	Interesting domain-specific data has been collected by a government official, yet the metadata describing the data is very limited and not sufficient for an outsider to make sense of the data.

(continued)

Table 4.1 (continued)

Types of organizational issues	Organizational issues	Example
Infrastructure and process-related issues	Lacking infrastructure and resources (including skills and training)	A municipality wants to become more transparent and show the municipality's inhabitants which data it collects, yet the municipality does not have the human and technical resources and infrastructure to make the data available to the public.
	Unclear or shared ownership	Two governmental organizations have worked together and integrated their data registers and datasets to obtain new insights. They share the ownership of the newly created dataset, but they disagree about opening the data.
	Changes to organizational processes required	A governmental organization willing to open data by default needs to change not only the data opening processes, but also the processes that precede the opening (e.g., during the data collection processes), since considerable metadata need to be collected simultaneously alongside the data itself. Changing work processes is complicated and may require additional work for several employees, whereas there are no direct incentives for them to change their work processes.
	Negative consequences for the government	Gas drillings in the Netherlands create large financial benefits for the government. Open data about earthquakes was used by lobbyists to demonstrate against the gas drillings that caused earthquakes in the northern part of the Netherlands. Under pressure, the Dutch government had to decide to reduce the amount of gas derived from this part of the Netherlands. Thus, the publication of government data resulted in less income from gas drillings.
	Benefits obtained by others than the government	The Ministry of Environment and Infrastructure puts much effort into opening datasets concerning traffic, road conditions, license plates and vehicle information. A company uses this data and creates an application that presents the information through a user-friendly interface that citizens need to pay for. The company creates revenue out of selling the application, whereas the government does not.

Adapted from Janssen et al. (2012), Susha et al. (2015), Zuiderwijk (2015a), Zuiderwijk and Janssen (2015), Zuiderwijk et al. (2012b)

4.2.1 Data-Related Issues

4.2.1.1 Potential Privacy Breaches

An important issue for governmental organizations opening data concerns the risk to violate individuals' privacy (Kalidien, Choenni, & Meijer, 2010; Kulk & van Loenen, 2012). Regardless of the amount of effort put into removing privacy sensitive content from datasets, privacy cannot be guaranteed. Even if an individual dataset does not violate a person's privacy, the combination of multiple datasets or the combination of open datasets with information from the media may allow for identifying persons in a dataset (Zuiderwijk & Janssen, 2014b), especially when open data is combined with social media data (Nieuwenhuijs, 2014). For instance, let us imagine that a researcher locates two datasets. The first dataset contains data about the number of crime offenders in a certain neighbourhood per type of crime (e.g., sex offences). With this dataset, someone can identify in which neighbourhood sex offenders live. The second dataset reveals the number of crime offenders per type of crime and per gender and age category. On themselves, these datasets do not allow identifying a particular person. However, their combination may allow this. If there is only one female sex offender in the age category of 70 years and older in a certain neighbourhood, identification of the particular offender becomes possible. With additional information from the media, the person might be identified. If one organisation releases the first dataset from the example and another organisation releases the second dataset, the privacy of citizens can easily be violated (example adapted from Kalidien et al., 2010).

Data protection legislation often prescribes on a very general level how one should handle privacy sensitive data, and thus it does not give much guidance for removing (privacy) sensitive information from datasets (Zuiderwijk & Janssen, 2014b). Laws and regulations need to give sufficient space for the interpretation of privacy sensitivity and therefore they cannot be too specific (idem). Furthermore, the situation in different countries might vary, as privacy is valued more in some countries than in others (idem). In sum, guidelines about privacy sensitivity partly help to identify which data cannot be published, yet much interpretation effort by the data provider is still required, and combining data could still lead to identifying a person or company (Zuiderwijk & Janssen, 2014b). When privacy-sensitive data is opened, this can result in considerable negative attention and might lead to reputation damage of the organization that opened the data or might lead to a decrease of trust in the government in general.

4.2.1.2 Data Sensitivity and Security

In addition to privacy-sensitive data, some governmental datasets are sensitive in other ways. For instance, data can be policy-sensitive. Whereas privacy-sensitive data refers to violating privacy of an individual or company, policy-sensitive data refers to data that may have negative consequences for government officials, responsible for a policy or for politicians working on issues related to these datasets. The data may contrast certain statements or positions, posited by a politician or it may show that a certain policy proposed by an important politician does not work as expected (Zuiderwijk, Janssen, Choenni, & Meijer, 2014). Governmental data may also be sensitive in the sense that it contains information that is considered a state secret and should not be provided to politicians of other countries, as it may block negotiation processes, or it may negatively influence ongoing alliances.

Sensitive data is often not released. Data sensitivity is an issue for organizations aiming to open up government data. On the one hand, these organizations are willing to become more open, yet on the other hand, determining whether a dataset is sensitive is complicated and accidentally releasing sensitive data could have many undesired consequences (Zuiderwijk et al., 2014). For example, opening sensitive data could damage the reputation of an individual (including politicians) or organization, it could also be dangerous, or lead to the resigning of a minister or conflicts with other countries.

Determining which data is sensitive and which data is not requires an examination of each individual dataset that an organization considers opening, also bearing in mind the context of to whom the data will be opened and with which other data the data might be released and potentially combined. This consideration requires interpretation by a human being, and mistakes might be made (Zuiderwijk, 2016). Since sensitive data is often not released, the data that is released usually favors policies set and arguments provided by politicians in place. Data that might demonstrate the opposite and give a different perspective might not be opened (Zuiderwijk & Janssen, 2014b).

4.2.1.3 Embargo Period

For each governmental dataset that is considered to be opened, a government official needs to ask the question of whether there are reasons for not yet opening the data, which may include an embargo period, i.e. a period in which a dataset is not publishable, although it might become publishable in the future. Some datasets may be publishable, but just not immediately after they have been collected. Reasons for having an embargo period are diverse. As an example, civil servants may first use the collected data to write a governmental report, and the opening of the data should be delayed until politicians presented and discussed the report (e.g., at the level of the national government) (Zuiderwijk & Janssen, 2014b). Some reports need to remain confidential, and thus the data also remains closed. A second reason for setting an embargo period is that civil servants may want to write an article (e.g., a scientific article) based on these data (idem). Data publication then has to be postponed, until the article has been published, which can take years. Other reasons for an embargo period include that the governmental organization that collected the data may want to conduct follow-up research using the data (idem), or that the data is too sensitive at a certain moment (e.g. when the national government is discussing a certain issue or topic and developing policies and legislation in this area). This data might become less sensitive over time.

Embargo periods have several advantages. Some datasets may still be opened when an embargo period is used, whereas they would not have been opened otherwise. Embargo periods give governmental organizations time to think data release through and may prevent wrongfully publishing data. It also allows for still publishing data that has become less sensitive over time. Embargo periods also have disadvantages. Datasets may become less useful over time; their quality reduces as timeliness of the data reduces at the moment of data publication (Zuiderwijk, 2016; Zuiderwijk & Janssen, 2014b).

4.2.1.4 Data Openness, Lack of Control Over Its Use and Lack of Trust in the Data User

To which extent should openness be provided? In one respect releasing governmental data may provide the public with more insight in what governmental processes encompass and what public agencies do. Datasets may be copied to many repositories and become available to a large audience. In another respect, opening governmental data to the public may result in too much openness. When datasets are open and become available at different places, this does not only enhance openness, but this also makes it difficult for the data provider to keep track of where the data is available and how it is used. The data provider may fear misuse of the data and may not completely trust potential data users. In addition, public agencies may accidentally release sensitive data that should not have been released. This may result in a more negative image of the government and may decrease the public's trust in the government.

4.2.1.5 Data Quality

Another consideration when opening governmental data concerns the quality of the data. Important data quality dimensions include completeness, timeliness, accuracy and consistency (Batini, Cappiello, Francalanci, & Maurino, 2009). Civil servants may decide to disclose data without having insight in its quality. Consequently, they may publish data that is incomplete, inaccurate, invalid, or unreliable. This may lead to low value and exploitation possibilities and thus to low reusability (also see Chap. 7 concerning value creation). Low quality data may also be published on purpose where publishing low quality data is considered a "quick win". Proponents for releasing and opening low data quality data argue that the release of low quality data could help in identifying the dimensions on which the quality of the data is poor, so that governmental data providers can improve these dimensions (see Chap. 7). The crowd can comment on the data and can try to improve low-quality data. Feedback to data providers regarding data quality might create incentives for the data publisher to improve the data (Zuiderwijk & Janssen, 2014b).

At the same time, some data users may not notice that the data is of poor quality. The low-quality data may be reused, and decisions and conclusions may be based on this data. This may result in wrongful decisions and little value creation. A dataset with many missing values or variables may be misinterpreted or may not be useful at all. Opponents of releasing low-quality data state that datasets need to have at least a certain level of quality before they can be published (Zuiderwijk, 2016) and should be in a format that enables reusability (also see Chap. 5 concerning interoperability). Both the arguments of the proponents and the opponents can be valid and assessing whether low-quality data can be opened requires a trade-off per dataset (Zuiderwijk & Janssen, 2014b). Data quality can also be subject of evaluation (see Chap. 8).

4.2.1.6 Data Documentation

Another consideration for releasing governmental data concerns data documentation. To be able to use open government data, users need to have information about the meaning of the data and the semantics need to be clear. They need data documentation to understand how the data can be used. For instance, to be able to find a book in the library, a person needs to know in which category he or she should look for the book. Part of the collected governmental data is poorly-documented and might be misinterpreted if it would be opened. Data may concern a specific domain (e.g. earth observations or the criminal justice chain) whereas data users do not necessarily have the domain-specific knowledge that is required to interpret the data correctly. This could lead to incorrect conclusions derived from data analysis results (Zuiderwijk & Janssen, 2014b). Considerable documentation is then required to understand the data. At the same time, adding considerable documentation to governmental datasets requires effort and time investments from the data provider, since this information often cannot be derived automatically from the data provider's systems (Zuiderwijk, 2016).

4.2.2 Infrastructure and Process-Related Issues

4.2.2.1 Lacking Infrastructure and Resources

Opening data requires the availability of an infrastructure. An Open Government Data Infrastructure can be defined as "*a shared, (quasi-)public, evolving system, consisting of a collection of interconnected social elements (e.g. user operations) and technical elements (e.g. open data analysis tools and technologies, open data services) which jointly allow for OGD use.*" (Zuiderwijk, 2015a, p. 45). Open data infrastructures are shared by a variety of actors and systems. Actors, such as governments, researchers, and citizens, can use the infrastructure, for example, by downloading and processing a dataset. Open data infrastructures consist of technical elements, such as tools and technologies (e.g., tools and platforms to analyze open data), and social elements, such as user operations and interactions (e.g., communication from the provider to the user about how the infrastructure can be used) (Zuiderwijk, 2017). Data, platforms and people are connected through the open data infrastructure (idem). Data, information, and knowledge are important resources that are transferred and exchanged in open data infrastructures. Such infrastructures evolve through the development of new technologies and through the adaptation of the infrastructure by people. All infrastructure elements are needed in combination to ensure that the infrastructure can function. The lacking or malfunctioning of one element results in problems for the functioning of the entire infrastructure. For example, if data providers and users are not connected, or if platforms are lacking of functionality and components, it becomes difficult to find and use the data and attain the potential benefits. In practice, open data infrastructures are still under development and various challenges need to be overcome. For instance, many open data infrastructures are mainly focused on the opening of governmental data and less on the use of the data, whereas the data use should eventually lead to attaining the benefits.

Opening data also requires resources of governmental organizations. Human resources are needed, such as computer skills, skills concerning data interpretation (to assess whether a dataset can be opened), resources for uploading datasets (e.g., time and effort), and resources related to the selection of tools for opening and sharing data. Data opening also requires technical resources, such as an internet connection and tools for processing and viewing datasets, as well as information and data resources, such as a repository of open data sets. Civil servants may need to be trained to develop the skills needed to open up governmental data.

4.2.2.2 Unclear or Shared Ownership

Data opening requires an assessment of ownership of the data. Often datasets are created through a collaboration of multiple people and organizations, and it may be unclear who owns the data, or involved parties may disagree about whether a dataset can be opened. Even if the collaborators agree on opening datasets that they created together, a potential risk is that it may be unclear who is responsible and accountable if something goes wrong, for instance, if data is misused. Datasets owned by organizations from different countries may also have to comply with different laws and policies concerning data protection (Faerman, McCaffrey, & Slyke, 2001; Zuiderwijk et al., 2014).

4.2.2.3 Changes to Organizational Processes Required

To really become open and systematically publish open datasets, governmental organizations need to make changes at different organizational layers (Van Veenstra & van den Broek, 2013) and for many organizations it is unclear how the publishing process could be modified to improve it and to institutionalize data opening

(Zuiderwijk & Janssen, 2013; Zuiderwijk et al., 2014). The open data literature is more focused on the development of open data portals and infrastructure, data publication, functionality and other instruments to release and use open data. Although this is an important first step, it is important to transform the structure of organizations and change the cultures and incentives to open data so that structural changes are made and so that opening data becomes part of the daily work processes, routines, and procedures (Zuiderwijk & Janssen, 2014b).

4.2.2.4 Negative Consequences for the Government

Releasing governmental data does not only have the potential to result in benefits, but can also lead to negative consequences for the government. Several scholars mention that opening data may result in, for example, the benefit of transparency (e.g., Bertot, Jaeger, & Grimes, 2010; Böhm et al., 2012a), yet transparency may also result in a more negative image of the government. If datasets of low quality are opened, or if opened datasets reveal the misbehavior of civil servants, this might decrease trust in the government (Zuiderwijk & Janssen, 2014b). Furthermore, opened datasets may be misused or misinterpreted (Kalidien et al., 2010; Kulk & van Loenen, 2012; Zuiderwijk et al., 2014).

4.2.2.5 Benefits Obtained by Others Than the Government

One of the challenging aspects of the open data process is that governmental organizations invest resources by opening data, whereas others benefit from this. The data providers are often not the ones who benefit, although they spend time and effort on opening the data. Policy makers working for governmental organizations may be able to use insights that data users outside the government obtained from the analysis of the governmental data. This may concern, for example, policy-making in the area of social security, economy, justice, elections, health, energy, and transport (Zuiderwijk, 2015a). Zuiderwijk (2015a, p. 4) describes the example of governmental policy-makers, who use insights obtained from the use of open crime data by non-governmental researchers to develop governmental policies about security measures and police surveillance. However, often users and (governmental) policy-makers do not communicate about the results of open data use and what lessons can be learned from this (Zuiderwijk, 2015a).

4.3 Use-Case: Solutions to Overcome the Issues

In this section, we discuss two use-cases that contain solutions on how to overcome some of the above-mentioned issues. They focus particularly on the risk of privacy violation (from an administrative perspective), and on the issue that benefits are usually obtained by others than the governmental organization that is opening the data (from a research perspective).

4.3.1 Solutions to Reduce the Risk of Privacy Violation (Administration View)

Yin (2017) provides an overview of solutions to enhance privacy and to reduce the risk on privacy violation for information sharing in general. He states that such solutions should combine technical and governance or managerial aspects. One category of technical aspects is referred to as Privacy Enhancing Technologies (PETs), including tools for encryption, policy, filtering, and anonymization (Yin, 2017). More examples concerning these PETs can be found in Seničar, Jerman-Blažič, and Klobučar (2003). The governance or managerial aspects mentioned by Yin (2017) include the development of legislation for data protection, self-regulation (voluntary privacy protection mechanisms) and privacy by design (building in privacy upfront). Privacy by design can be defined as an approach to protect privacy by embedding it into the design specifications of technologies, business practices, and physical infrastructures (Cavoukian, 2011).

In addition, Ali- Eldin, Zuiderwijk, and Janssen (2017) developed a model for privacy risk scoring for open data. The model consists of open data attributes and privacy risk mitigation measures. The open data attributes influence the decision of whether or not to open up a dataset. They include the need for openness, the criticality/importance level, the level of cyber security threat, the trustworthiness of the data provider, and the restrictions of use (including the type of user, the physical location that the data is accessed from, and the purpose the data is used for). Each attribute has different values and each value has a different score. Adding up the scores results in a Privacy Risk Indicator (low, low-medium, medium-high, or high). Based on the indicator level, a Privacy Risk Mitigation Measure (PRMM) is proposed. If the PRI is low, only the removal of identifiers from a dataset is proposed, using tools such as Anonymizer, ARX, or Camouflage's-CX-Mask. If the IRP is on the low-medium level, the model recommends altering quasi-identifiers to reduce identity leakage. "Quasi-identifiers are data types which if linked with other datasets can reveal real identities" (p. 150). If the IRP indicates medium-high privacy risks, the model suggests removing sensitive items, and when the IRP is high, it is advised not to publish the data at all. Each defined privacy risk mitigation measure should be applied before publishing a government dataset on the internet (Ali-Eldin et al., 2017).

These are just a few examples of data protection solutions, but more of them exist. Furthermore, each of the provided solutions also has its drawbacks. For instance, anonymization is often not sufficient, as the combinations of datasets could still lead to re-identifying persons and their activities.

4.3.2 Solutions to Develop an Open Data Infrastructure That Enhances the Coordination Between Open Data Actors (Research View)

In practice, benefits of open data are usually obtained by others than the governmental organization that is opening the data. Zuiderwijk (2015a) argues that the use of open government can support open data publication and governmental policy-making, since governmental open data providers and governmental policy makers can learn from the insights obtained using open data. This is challenging, since this requires several actors – dependent on each other – to work together and to coordinate their activities. Zuiderwijk (2015a) proposes the design of an open data infrastructure to enhance the coordination of open data use by researchers. An infrastructure for Open Government Data (OGD) is defined as a shared, (quasi-) public, evolving system, consisting of a collection of interconnected social elements (e.g., user operations) and technical elements (e.g., open data analysis tools and technologies, open data services) which jointly allow for OGD use (p. 269). The theory focuses on the coordination of searching for and finding OGD, OGD analysis, OGD visualization, interaction about OGD, and OGD quality analysis. "*In the context of this study, three design propositions were elicited:*

- Metadata positively influence the ease and speed of searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD and OGD quality analysis.
- Interaction mechanisms positively influence the ease and speed of interaction about OGD.
- Data quality indicators positively influence the ease and speed of OGD quality analysis." (Zuiderwijk, 2015a, p. 270)

The metadata model, the interaction mechanisms, and the data quality indicators need to be combined to support searching for and finding OGD, OGD analysis, OGD visualisation, interaction about OGD, and OGD quality analysis. Building on 22 coordination design principles, 40 metadata design principles, 15 interaction design principles, and 4 data quality design principles, the system design, the coordination patterns and the function design of the OGD infrastructure were developed. Evaluations of a prototype, integrating the designed infrastructure, provided support for the three propositions (Zuiderwijk, 2015a).

4.4 Best Practices

The Share-PSI 2.0 project has created an overview of best practices for sharing open government data (Share-PSI 2.0, 2016a), as depicted in Table 4.2. One of the main aims of the Share PSI 2.0 best practices is the Implementation of the (revised) PSI Directive (European Commission, 2003, 2013c).

Table 4.2 Best practices for sharing open government data

Best practice (Share-PSI 2.0, 2016a)	Description (Share-PSI 2.0, 2016a)
Categorise openness of data	Public sector organizations can create a system in which the openness of data is categorized, so that it becomes easier for them to determine with whom data can be shared.
Dataset criteria	Public sector organizations can prioritize the publication of some datasets in comparison to others. For example, datasets that contribute to transparency, datasets that help with cost reductions, or highly structured datasets may be published first.
Develop an Open Data publication plan	Public sector organizations are recommended to develop a plan in which they address the abovementioned issues and determine which datasets are fit for publication as open data, which requirements the internal and external stakeholders have, and which potential benefits, risks and costs of data opening play a role.
Develop and implement a cross agency strategy	In addition to a plan for individual organizations, it is recommended to develop and implement an open data strategy that coordinates the efforts of multiple organizations. In most of the EU countries these strategies have been interpreted in guides for publishing data across agencies and in some cases, they are incorporated in the national law through presidential or ministerial decrees. The strategy should also foresee the way the opening will be implemented by the public-sector organisations. An example could be a stage strategy focusing at the first level in a quick win publishing data as quickly as possible before a specific deadline and at the second level focusing on the quality improvement (Share-PSI 2.0, 2016b).
Enable feedback channels for improving the quality of existing government data	The quality of governmental data can be improved by facilitating feedback channels for users to report errors, inconsistencies, and incompleteness in openly available datasets.
Enable quality assessment of open data	Since data quality is considered to be subjective, depending on the context, data quality should be measured in different ways all along the data pipeline (not only at the front end). These measures should sustainably raise data quality.
Encourage crowdsourcing around PSI	The open data community can help to improve the quality and quantity of available datasets and can enthuse potential data users.
Establish an Open Data ecosystem	An open data ecosystem can enable the uptake of government data and information for reuse, so that services can be built for citizens.
Establish Open Government Portal for data sharing	Government data should be published through open data portals that provide potential users with easy access to a searchable hub for multiple datasets.
High level support	Senior staff should support open data actions.
Holistic metrics	Value generation using open government data and costs of making this data available have to be assessed in respect to large-scale detour effects and not only at the level of the data providing agency.

(continued)

Table 4.2 (continued)

Best practice (Share-PSI 2.0, 2016a)	Description (Share-PSI 2.0, 2016a)
Identify what you already publish	To make it easier to decide what data should be made available, it is useful to examine which datasets are already opened. An inventory must be created and maintained of already opened data.
Open Data business models and value disciplines	A business model should be described, explaining how value is created and captured for data opened by a certain public organization (at all levels) and what the expected results are.
Open up public transport data	Transport data (e.g. timetables, service disruptions and accessibility) is considered as high-value data and can be used to create a better experience for transport users, greener cities by using collective transport, and more efficient companies.
Open up research data	Opening up research data promotes the discoverability and measurability of scientific achievements, and can stimulate innovation, economic growth and education.
Provide PSI at zero charge	The ability to use open data without payment unlocks maximum commercial and non-commercial potential.
Publish overview of managed data	Public organizations must publish an overview of datasets that it manages, so that potential users know what may be(come) available.
Publish statistical data in Linked Data format	The Linked Data format is an approach for expressing data in a standardised machine-readable manner and for providing a recommended set of metadata terms to describe the data.
(Re)use federated tools	Federated/distributed tools for open data collection can be used to automatically publish all the (meta)data published on the websites of each public entity. This can result in a global index of reusable open datasets.
Standards for Geospatial Data	For many public and privacy organizations location is essential and thus geospatial data should be shared in a way most likely to be re-usable: adhering to standards.
Support Open Data start ups	Open data provides a good basis for entrepreneurship, allowing for the development of added value services by citizens and small enterprises. Start-ups can be supported through the collaboration between universities (potential entrepreneurs), private and public funding organisations (chambers of commerce, municipalities, start-up investors) and experts (coaches and mentors).

Share-PSI 2.0 (2016a)

In addition, technical best practices related to the publication and usage of data on the Web have been developed by the World Wide Web Consortium (W3C) (World Wide Web Consortium, 2017). The best practices facilitate the interaction between data publishers and data users, and emphasizes that data should be discoverable and understandable by humans and machines. It also states that the use of data should be discoverable and that the efforts of the data publisher should be acknowledged and recognized (Table 4.3).

More information concerning each W3C Best Practice can be found at http://www.w3.org/TR/dwbp/.

Table 4.3 Technical best practices related to the publication and usage of data on the Web

	Best practice (World Wide Web Consortium, 2017)	Description (World Wide Web Consortium, 2017)
Metadata	1. Provide metadata	Provide metadata for both human users and computer applications.
	2. Provide descriptive metadata	Provide metadata that describes the overall features of datasets and distributions.
	3. Provide structural metadata	Provide metadata that describes the schema and internal structure of a distribution.
Data licenses	4. Provide data license information	Provide a link to or copy of the license agreement that controls use of the data.
Data provenance	5. Provide data provenance information	Provide complete information about the origins of the data and any changes you have made.
Data quality	6. Provide data quality information	Provide information about data quality and fitness for particular purposes.
Data versioning	7. Provide a version indicator	Assign and indicate a version number or date for each dataset.
	8. Provide version history	Provide a complete version history that explains the changes made in each version.
Data identifiers	9. Use persistent URIs as identifiers of datasets	Identify each dataset by a carefully chosen, persistent URI.
	10. Use persistent URIs as identifiers within datasets	Reuse other people's URIs as identifiers within datasets where possible.
	11. Assign URIs to dataset versions and series	Assign URIs to individual versions of datasets as well as to the overall series.
Data formats	12. Use machine-readable standardized data formats	Make data available in a machine-readable, standardized data format that is well suited to its intended or potential use.
	13. Use locale-neutral data representations	Use locale-neutral data structures and values, or, where that is not possible, provide metadata about the locale used by data values.
	14. Provide data in multiple formats	Make data available in multiple formats when more than one format suits its intended or potential use.
Data vocabularies	15. Reuse vocabularies, preferably standardized ones	Use terms from shared vocabularies, preferably standardized ones, to encode data and metadata.
	16. Choose the right formalization level	Opt for a level of formal semantics that fits both data and the most-likely applications.

(continued)

Table 4.3 (continued)

	Best practice (World Wide Web Consortium, 2017)	Description (World Wide Web Consortium, 2017)
Data access	17. Provide bulk download	Enable consumers to retrieve the full dataset with a single request.
	18. Provide Subsets for Large Datasets	If your dataset is large, enable users and applications to readily work with useful subsets of your data.
	19. Use content negotiation for serving data available in multiple formats	Use content negotiation in addition to file extensions for serving data available in multiple formats.
	20. Provide real-time access	When data is produced in real-time, make it available on the web in real-time or near real-time.
	21. Provide data up to date	Make data available in an up-to-date manner, and make the update frequency explicit.
	22. Provide an explanation for data that is not available	For data that is not available, provide an explanation about how the data can be accessed and who can access it.
Data access – APIs	23. Make data available through an API	Offer an API to serve data, if you have the resources to do so.
	24. Use Web Standards as the foundation of APIs	When designing APIs, use an architectural style that is founded on the technologies of the web itself.
	25. Provide complete documentation for your API	Provide complete information on the web about your API. Update documentation as you add features or make changes.
	26. Avoid Breaking Changes to Your API	Avoid changes to your API that break client code, and communicate any changes in your API to your developers when evolution happens.
Data preservation	27. Preserve identifiers	When removing data from the web, preserve the identifier and provide information about the archived resource.
	28. Assess dataset coverage	Assess the coverage of a dataset prior to its preservation.
Feedback	29. Gather feedback from data consumers	Provide a readily discoverable means for consumers to offer feedback.
	30. Make feedback available	Make consumer feedback about datasets and distributions publicly available.
Data enrichment	31. Enrich data by generating new data	Enrich your data by generating new data when doing so will enhance its value.
	32. Provide Complementary Presentations	Enrich data by presenting it in complementary, immediately informative ways, such as visualizations, tables, web applications, or summaries.

(continued)

Table 4.3 (continued)

	Best practice (World Wide Web Consortium, 2017)	Description (World Wide Web Consortium, 2017)
Republication	33. Provide Feedback to the Original Publisher	Let the original publisher know when you are reusing their data. If you find an error or have suggestions or compliments, let them know.
	34. Follow Licensing Terms	Find and follow the licensing requirements from the original publisher of the dataset.
	35. Cite the Original Publication	Acknowledge the source of your data in metadata. If you provide a user interface, include the citation visibly in the interface.

World Wide Web Consortium (2017)

4.5 Conclusions

In sum, opening government data is not easy, and there are many aspects that need to be considered when a public agency decides to open datasets. In this chapter we identified 11 organizational issues for opening up government data. These encompass six data-related issues (potential privacy breaches, data sensitivity and security, embargo period, data openness, lack of control over its usage and lack of trust in the data user, data quality, and data documentation) and five infrastructure and process-related issues (lacking infrastructure and resources, unclear or shared ownership, changes to organizational processes required, negative consequences for the government, and benefits obtained by others than the government).

When governments consider opening their data, they need to make a trade-off between the potential benefits and the potential disadvantages of this decision. A key question is: to open or not to open the data? The data requires a trade-off in which either the benefits or risks of opening may dominate. Figure 4.1 shows the decision-making process in which the benefits and disadvantages of opening data are weighed. Some data has many benefits and hardly any disadvantages and can be opened without any discussion. Other data should not be opened without any doubt due to security, privacy, or other reasons. There is a huge pile of data requiring a trade-off in which either the benefits or risks may dominate.

We do not know how large this part is that organizations need to decide on. Furthermore, it is likely that this changes over time. Since public values represent the needs and preferences of the collective citizenry, public values may change over time, as the needs and preferences of citizens may also change. It is likely that the decision regarding which data should be opened or closed will vary over time.

Thus, the most important trade-off is to open or not to open the data. This trade-off is based on the considerations that we described, such as data quality and data sensitivity. For each of the considerations, the civil servant responsible for data release needs to decide which aspects are more important. For instance, is it more

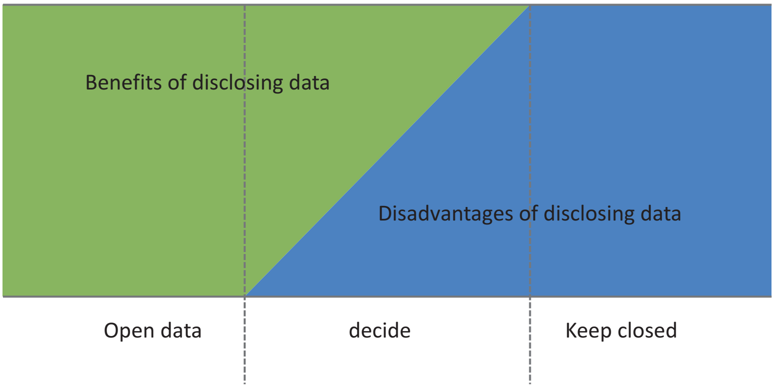

Fig. 4.1 Decision-making to open or not to open datasets (Zuiderwijk & Janssen, 2015, p. 114)

important that data are of high quality or is it more important just to publish the data and to let data users point out aspects of low quality? Is it more important to ensure that absolutely no datasets are published which are sensitive, and to remove all potentially sensitive variables? Or is it more important that the data is more useful, but might potentially be sensitive when combined with other data?

This chapter also provided several use-cases that describe how some of the identified issues can be overcome. The use-cases focused on solutions to reduce the risk of privacy violation (from an administration view) and on solutions to develop an open data infrastructure that enhances the coordination between open data actors (from a research view). Furthermore, we examined best practices as provided by the PSI-Share project and by the World Wide Web Consortium. Following these best practices should make it easier to reap the benefits of open data, as described in Chap. 1 of this book.