D4.5 Legal and privacy toolkit v2

Coordinator: Walter Palmetshofer, OKF

With contributions from: Julia Manske, Tobias Knobloch, Michael Peters, OKF

Reviewer: Elena Simperl, SOTON

Deliverable nature	Report (R)
Dissemination level (Confidentiality)	Public (PU)
Contractual delivery date	31th July 2017
Actual delivery date	9th August 2017
Keywords	personal data, data protection, risk assessment

Executive summary

Open government data is nowadays an integral part of the digitalization strategy of most European governments. They are a vital part of the data ecosystem and a source for innovative solutions. Open data is used by the administrations itself, by civil society, start-ups and established companies and research department.

Since the 15th of July 2017 France, United Kingdom and Germany finally have open data laws in effect. With this, an important step has been taken. The next challenge is to open up government and business data in responsible manner. In Germany there is a high degree of privacy and therefore a strong public interest in its protection. In a time of continuous data processing, automated processing, and the spread of data-driven business models, a purely legal view of data protection is not sufficient to ensure the protection of privacy.

However, there are solutions for this problem. In the light of the new open data law and its impending implementation, this toolkit provides six general recommendations:

1. More than a box ticking exercise - capacity building and resources
2. Risk analysis of the data
3. It does not work without high quality technical data protection
4. Perform regular risk assessments
5. Consideration of regulatory approaches
6. Cross-linking of experts and expertise

Subsequently, in the description of the toolbox concrete suggestions are made on how to organize processes before, both during and after data opening, and which tools can be used to adequately address the data protection risks.

The report draws on concrete cases to illustrate data evaluation methods and tools. The case of San Francisco[1] traffic light system provides important lessons for the improvement of the quality of anonymisation procedures when the data is opened. Secondly, it is recommended that employees are trained in anonymisation procedures, accessing technical support, and recording metadata on the type of anonymisation method applied. Thirdly, the access systems to specific highly valuable datasets must be restrictive, for their inherent sensitivity. Finally, continuous “stress tests” are recommended, these should be carried out by external experts, who would have to examine the risk of de-anonymisation of data records with the help of other data sets.

This main example and the description of the appropriate instruments is complemented with examples from abroad and other fields of work. Finally, the report discusses the chances and risks when opening data and provides a repertoire of tools for employees to open data and reduce risks simultaneously. This report complements D4.4 Legal and privacy toolkit v1 (and the aim of that deliverable of “What are the critical things to consider when opening up data?” with the focus on privacy of the data based on the demand by ODINE startup applicants) with a long term view and procedures.

[1] https://datasf.org/blog/4-steps-to-manage-privacy-and-de-identification-for-your-open-data-program/

Context and six basic recommendations

Open government data is nowadays an integral part of the digitalization strategy of some European governments. They are a vital part of the data ecosystem and a source for innovation. Open data is used by the administrations itself, by civil society, start-ups and established companies and research department.
Some open data sources become even more interesting when it is combined with other data sources. It also follows that, just as data from other sources, for example from the private sector, they are becoming part of the big data cosmos and contributing to the challenges currently being discussed in this context. In the first place, this concerns the protection of privacy. As current examples show, the societal implications of data usage can go far beyond classical data protection, for example if discrimination is intensified.[2] International examples show that open administrative data can also contribute to these problems.[3]

Challenges associated with the opening of administrative data are classified as follows:[4]

- Data sets are opened or excluded from the opening for incorrect or non-transparent reasons[5]
- Data is being opened incorrectly, for example by using poor quality anonymisation methods
- Anonymized data can be de-anonymized by linking to other data sources, which may also be publicly available
- Open administrative data are used for purposes that are unlawful or unethical by the public

Opening up data will hopefully further gain more traction, not just open government data but also open data from businesses. Therefore, it will be important to create the right framework in the beginning in order to prevent restrictions of the basic rights of citizens or open data being misused (for example profile creation). Independent of the legal dimension, citizens' and consumer trust in the long-term success of open data will be decisive.

For that trust building we provide six recommendations[6].

[2] As example O'Neil, C. (2016): Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, New York

[3] Manske, J. (2016). Offene Daten und der Schutz der Privatsphäre. https://www.stiftung-nv.de/de/publikation/offene-daten-und-der-schutz-der-privatsphaere.

[4] Based on research, working groups and collected feedback from open data experts.

[5] We would like to emphasize once again that we are specifically focusing on the protection of privacy and informational self-determination among the citizens in Germany. Data protection, on the other hand, must not be used as a pretext to make data - which for example serve the transparency of state action - not made publicly accessible. This has happened in the past in the context of the Information Freedom Act.

[6] Based on research, working groups and collected feedback from open data experts.

Firstly, a introduction of clear and transparent standards and principles for an assessment, whether data should be opened. This is also so important in order to prevent the use of data protection as excuse to stop open up relevant data.
Secondly, the requirements of the European General Data Protection Basic Regulation (GDPR) must be taken into account when implementing an open data strategy.
Thirdly, standards and capacity building for the anonymity of records are required.
Fourthly, it should not be ignored that research clearly shows the limits of the anonymisation of data, so that this measure alone will not suffice to ensure the protection of privacy.
Fifthly, open data must be understood as part of the general discourse on responsible data use, and in so far as it is taken into account when considering the opportunities and risks of data in general.

These general recommendations should now be supported by concrete processes and instruments in this privacy toolkit version 2.0. Many of the listed approaches are gathered from around the world and inspired by procedures from similar areas, such as IT security.

A first version of this collection of instruments was evaluated in a workshop in November 2016[7] with experts from administration, data protection, IT and civil society as well as in individual discussions with further experts. Afterwards, we discussed it with open data practitioners[8] to see if this original approach focusing on open government data is also valid for general open data (including for businesses).

The approaches are based on three process phases of open data:

1. Before publication
2. When publishing
3. After publication

The presentation of the concrete instruments with the discussion of the respective opportunities and risks as well as references to examples from other areas or countries can be found in Section 2.

The description of the process sequence at the beginning of the three sketched phases should specify how the implementation of the instruments could look in practice.
On the basis of our research and intensive dialogues at the interface open data and data protection, we would also like to make the following six basic recommendations for a strong consideration of data protection aspects in the provision of open data.

1. More than a box ticking exercise - capacity building and resources

We would like to encourage administrative staff to provide a holistic perspective on the benefits and risks of open data. As long as data protection is only considered as a legal

[7] In Berlin 2016 at SNV https://www.stiftung-nv.de/de/veranstaltung/open-data-privacy-workshop
[8] From Deutsche Bahn, Bitkom, ODINE startups and community

requirement which has to be checked off but not as a strategically integrated processes, there will be a risk that data opening is more likely to cause social problems rather than solve them. This is less expensive than it may sound, after all, the administration has to build capacity for open data anyway. For this, resources are required. We advocate that this should also take into account the development of capacities for the protection of privacy and be included in resource planning. From our point of view, it would be a missed opportunity if no competences for the risk analysis of data as well as for the technical data protection were located in such an office. After all, special knowledge is required in order to recognize and assess both opportunities and risks adequately (the generalist principle here comes to a limit). As soon as possible, guides and checklists should be provided by this office for finding, classifying, preparing and publishing data sets.

These materials represent an important first and above all low-threshold support for employees. However, they will not be sufficient to assess the potential risks of data in its entirety. Rather administrative staff must be educated in professional trainings (for example in data-competency, like GDS - Government Digital Service in UK demonstrates[9], as well as via (Blended Learning, e-learning-offerings). An open-data-friendly infrastructure also favors the protection of data by simplifying the control and standardization of processes around the data, thereby providing a better overview of the data. This in turn makes it easier to find or avoid errors.

2. Risk analysis of the data

The better the knowledge about opening already opened data is, the easier it is to estimate the possible risk from them. To this extent data sets should be categorized by default.[10]
The checklists mentioned above could be used to categorize data using a traffic light system.
"Red data" should generally not be opened at all, but "green", on the other hand, which clearly and most likely do not hold a any personal information even in the future, could be provided without any problems. The classification should be done at the data collecting, storage and processing centers, as they know the data best. The four-eyes principle must be applied. Once established, such a categorization can lead to a quick decision about the opening of data.

A special focus, and maybe potential additional effort, would be the examination of the "orange data". At the latest for these data records, the person responsible for the subject must be supplemented by a data protection expert.[11] Such a function could, for example, be exercised by an Open Data Advisory Board. It should be noted that thematic-specific, application-specific approaches will be indispensable since the data sets of certain work areas (geo data, health data, …) tend to carry a higher risk. Although the assessment of individual data sets is important in itself, this is not sufficient for a sound assessment of possible risks. For this purpose, the singular data must always be considered in conjunction

[9] https://gds.blog.gov.uk/2016/04/27/data-literacy-helping-non-data-specialists-make-the-most-of-data-science
[10] A catalogue of the data sets would also be helpful in the estimation of data protection risks. However such an approach can be resources intensive in the too long-term and difficult to implement.
[11] To our surprise, a number of administrative staff members even spoke about the establishment of a mediating agency with the appropriate expertise.

with other data sets. Moreover, once again, the risks arising from data sets are dynamic. This means that every categorization must also be subjected to regular checks (For the aspect of continuous checking please see 4. Perform regular risk assessments). On the one hand, because the risk increases with the data, and on the other hand because the definition of what is socially relevant or sensitive is subject to continuous negotiation processes.

This is illustrated by the increasing publication of previously hidden public transport data on the one the hand and, on the other hand, the reluctance to publish data on refugee facilities or critical infrastructure data such as power grid networks.

3. It does not work without high quality technical data protection

High-quality technical data protection procedures play a central role in the responsible opening-up of data, especially if we are talking about a broad and wide data supply.
The current discrepancy between the technical and legal evaluation of data protection is serious. In Germany the new open data law explicitly supports the opening up of all data, which by definition are not personal data and which must not be protected by other reasons (eg secret protection, FOIA, ...).
This means that anonymous data records - that is, those in which personal references were previously contained and then removed - should be opened. A lawyer would therefore always agree to the publication of anonymous data records. However, the legality is not yet a guarantor for a responsible handling of the thus opened data. And anonymisation is not always the same, even if it may seem so. The extent of anonymisation and the level of security attained thereby can be very different. In this context, it is to be regretted that no statement is made in the open data law as to which authority is responsible for the anonymisation of data records. The problem with the situation in Germany is that different standards apply to anonymisation, for example regarding the aggregation level of data. Municipalities, companies and countries assess the data protection risk differently.
It would be desirable to harmonize the legal interpretations. In any case, a high-quality anonymisation of data is very difficult considering the fast, automated processing of large amounts of data.Numerous research even points out that a complete and permanently effective anonymisation according to the latest state of technology can not be guaranteed at all.[12] One can therefore at best speak of an approach to a high level of data protection, but not of a guarantor. Ignoring these findings is not an option in the development and implementation of open data strategies of the different administrative levels.

This does not mean that anonymisation should be abandoned as a procedure.
Rather, as far as it is complete and of high quality, it is to be regarded as a necessary, but not sufficient element. In order to achieve a high quality standard, appropriate structures have to be created. The effort required for this should not to be underestimated, as already

[12] Ohm, P. (2009/2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. In: UCLA Law Review, 57, S. 1701–1777, U of Colorado Law Legal Studies Research Paper 9–12. http://ssrn.com/abstract=1450006; de Montjoye Y.-A.; Radaelli, L.; Singh, V. K.; Pentland, A. S. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. In: Science, 347 (6221), S. 536–539. DOI:10.1126/science.1256297.

mentioned capacity building within the administration is required. In order to meet the complexity of today's anonymisation procedures, administrative staff has to be trained. On the other hand, if a good data protection is actually a concern, not only the legal view but also a fundamental technical view of data must always be taken into account.

Up to now, there has been a rather negative attitude towards the use of technical data protection. Of course it can not be a solution solely relying on technical procedures.

Corresponding supporting tools could, however, significantly facilitate the work of the authorities and minimize errors; especially if they were to be integrated in specialist procedures. Previous software options may not yet be eligible. It would be desirable to invest in the development of user-friendly applications based on existing technologies specifically for the provision of open data by authorities and companies.

4. Perform regular risk assessments

As in all other areas where data is being handled, data protection must be understood as a process rather than a one-time audit. This must of course also be reflected in budgets and resource planning. Due to rapid technological changes, the risk of violating privacy can increase significantly within a short period of time. At the same time, however, it is also possible to develop new technical solutions that better protect data. If nothing else the GDPR will promote research and development in the field of data protection-friendly technologies (PET).[13]

To this extent, potential risks of data protection must be checked on an ongoing basis. In the case of such "anonymisation stress tests", the up-to-dateness and quality of the anonymisation procedures should be assessed.It is then necessary to focus on the entire data ecosystem and not on singular data sets.

Guiding questions for a corresponding audit would be:

- What are the intersections between the data records on the data platform and what conclusions can be drawn?
- What are the intersections of open data from other platforms? (open gov data, federal, regional level and open business data)
- Which data can be easily deanonymized by adding other (eg commercial, ...) data records?

In a sample-like manner, a regular check should be carried out to determine whether and why there is a risk of deanonymization. We are encouraging to promote the development of

[13] Report of European Union Agency for Network and Information Security, ENISA (2015). Privacy by design in big data https://www.enisa.europa.eu/publications/big-data-protection

appropriate technical support instruments and to examine whether and to what extent they can be integrated into specialist applications.
In the course of the legally required open-data capability of new information processing systems, the technical data protection aspect must be adequately considered.

At regular intervals a risk assessment by an external auditor with correspondingly developed technical expertise should be assigned. It is here to discuss who should take on this task. The scope of the independent data protection authorities, which are known to be already resource-poor, is normally limited to consulting.
However, established procedures in the field of IT security could provide an guidance. There, external hackers conduct planned penetration tests to check the security of the systems (also called "bug bounty programs").

In order to be able to react more dynamically to mistakes and to learn from them in time, it helps to consider and evaluate cases from home and abroad. Of course, many examples from other countries are always transferable because of different legal frameworks. Nonetheless a systematic analysis of mistakes or misuse cases (ideally including the unhinged anonymisation) can provide a basis for the evaluation of data (see the traffic light system) and increase the competencies within the administration.

We also advocate greater transparency on the nature of the anonymisation procedures used. The used anonymisation methods should be pointed out on the data platforms themselves. Another measure that would be particularly welcome from the point of view of technical data protection would be the documentation and publication of the anonymisation method used in the metadata for the respective data set.

5. Consideration of regulatory approaches

Even if there is a lot of reasons against regulatory approaches we would not like to rule them out categorically.

We are increasingly aware of the fact that data is being misused for cyber crimes or used for business practices which are unacceptable from a public perspective (For example when targeted loans are provided to socially disadvantaged people).[14] Of course, most of their data comes from commercial sources, but the Federal Trade Commission report from the US (there is no comparable information available for Europe) shows at least that open data is one of the information sources used by data dealers. More and more cases show that anonymised data is also de-anonymized and reused. Anonymized data can, for example, also be used to carry out specific fraud attacks on persons. Of course open data only has a small share in the general problem of the possible violation of civil rights by analyzing data. The Open Data movement will not solve this problem.

In view of the fact, that opening data is an idea which is oriented towards the common good,

[14] Online Lead Generation and Payday Loans. https://www.teamupturn.com/reports/2015/led-astray

everyone involved in that area should be willing to work on solutions and to commit to a culture of appropriate data usage. Regulation is one of the possible instruments, which should not be categorically excluded prematurely.

Although we consider the approach to declare the act of deanonymization as a crime, as currently implemented in Australia and considered in New Zealand[15], critical.
However, there are other ways to eliminate abuse, such as sanctioning trade, the use of deanonymized data or the use of data for certain "non-common-good-oriented" purposes.

If we are interested in opportunities and right-based data usage, we need more differentiated approaches both for the data access and the data use. Open data as a topic is a good opportunity to begin to think about how each user group gets which types of data usage.

Naturally the dynamics of global data flows make enforcement of national laws more difficult, but the limitation of the legal use of data has also proven to be effective in other areas[16].
For example, in archival legislation, the use of personal information is permitted for research, with the proviso that it can not be published.[17] Regulatory approaches are also used in statistical law.[18]

And even the existing data protection law in Germany restricts the use of data in any case in the sense that, in the case of public data, further use is permitted only if the legitimate interest of possibly affected persons is not disproportionately affected.

For the future, there are also research approaches at a technical level which are devoted to the development of so-called "sticky policies".[19] With these, records would be invariably marked so that any unlawful use would be traceable in the long term.

Also limited access, e.g. for certified researchers, to specific data sets in connection with licenses or controllable interfaces[20] are promising.

---

[15] In Australia since the amended Privacy Act (Re-identification Offence Bill, 2016) the reidentification of public data is a criminal offense http://www.lexology.com/library/detail.aspx?g=0e2a052e-d347-4421-8525-7c9c7b3c6dc4
Report to the Minister of Justice under Section 26 of the Privacy Act. Six Recommendations for Privacy Act Reform.
https://privacy.org.nz/assets/Files/Reports-to-ParlGovt/OPC-report-to-the-Minister-of-Justice-under-Section-26-of-the-Privacy-Act.pdf

[16] Although this should generally not be an exclusion criterion, this generally applies to the use of data. After all the GDPR will already harmonize the European market.

[17] Many thanks to Dr. Alexander Dix for this suggestion.

[18] See, for example, Section 21 of the Federal Statistics Act on the Prohibition of Reidentification. Dix, A. (2016). Datenschutz im Zeitalter von Big Data. Wie steht es um den Schutz der Privatsphäre? In: Stadtforschung und Statistik, 1, page 59–64. https://www.eaid-berlin.de/wp-content/uploads/2016/05/StSt-1-2016_Dix.pdf

[19] Pearson, S.; Casassa Mont, M. (2011). Sticky Policies: An Approach for Managing Privacy across Multiple Parties. https://documents.epfl.ch/users/a/ay/ayday/www/mini_project/Sticky%20Policies.pdf

[20] The Fraunhofer IESE has developed a ready to use technical application for data usage control in the industrial sector, see IND²UCE (Integrated Distributed Data Usage Control Enforcement) Framework https://www.iese.fraunhofer.de/de/competencies/security/ind2uce-framework.html

6. Cross-linking of experts and expertise

Once again, the networking of the expertise of several disciplines is necessary.
As explained here, in our view, it is not enough to look solely on the legal aspect of data usage. Although this view is indisputably essential even from a compliance perspective. However, in order to pursue the goal of a common-good-oriented and responsible opening of data, we urgently need to broaden this perspective.
In the first, place we see the stronger integration of technical data protection experts. Secondly, exchanges between data protectors (technical and legal), the open data community and the administration must be intensified.
This competence core is usually supplemented by expertise from areas that could benefit from or be adversely affected by open data (for example, care for the elderly, refugee care, AIDS counseling, etc.).

Toolbox

This toolbox is for protecting the privacy when opening data.

1. Definition of open by default

Open by default[21] means that the machine-readable-format data can be used freely, modified, and shared, standardized by anyone for any purpose.[22]

2. Exceptions of open data by default

2.1 Exceptions of open data by default due the legal framework

Exceptions of open data by default are depending on the legal situation (varies from country to country)

- Data protection laws
- Freedom of Information laws
- Further special laws which are limiting like national security
- Trade and business secrets
- Further secret protection (tax law protection in Germany, professional secrecy doctors, lawyers, …)
- Protection of intellectual property like copyright law, trademark law, patent law

[21] See open definition http://opendefinition.org/od/2.1/en/
[22] Ideal case for open government data, not for businesses. See the ODI data spectrum https://theodi.org/data-spectrum for further infos for businesses.

Furthermore, the interest in the publication must always be weighed against the following legal goods to be protected:

- Any personal or individual-related data
- Public concerns and legal enforcement
- Course of administrative procedures and upcoming administrative measures
- Protection of the decision-making process

2.2 Exceptions of the exceptions

The exceptions listed in 2.1 are the framework, but not the final decision. The framework leaves the possibility to publish data, either due to the public interest (important is here, that there a guidelines and the decision is not based on subjective judgement by a single person) or that the data can be so transformed, that the still can be published.

For example individual-related data is by definition excluded from publication, but anonymised data, can be published.

2.3 The scope of the to be published data must be further restricted to protect privacy

Even if legally possible by the framework in 2.1 not all data can be published as open data without any risk even if the data is anonymized.
The reasons therefor are

- technical limitations of anonymising procedures
- new risks due to combination with other data sets

For this reason, appropriate tools are needed to estimate the possible privacy risks before the publication and to minimize the risk during publication. The suggestions and ideas in this overview serve this purpose.

3. Proposed a three-phase risk assessment model for Open Data[23]

In the following, we differentiate these three process phases of the data opening and propose data protection measures for the respective phase:

1. ***Before publication****:* Decision whether data can or should be open at all
2. ***When publishing:*** Data protection measures in the course of the publication of data
3. ***After publication:*** Control of the use of already opened data

All measures are, of course, based on ongoing business processes, in which data protection principles - such as data protection and data thriftiness[24], for example in the procurement of IT systems - are already anchored.

[23] According to SNV and SF xxx

Another view at the different phases of open data could divided into the following stages:

- Collecting data
- Maintaining data
- Releasing data
- Deleting data

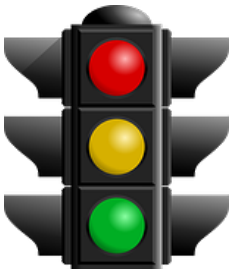
The different types of measures are distinguished and named as follows:

- Evaluation Assistance: Assistance to evaluate the data protection risk of one or more data records
- Institutionalization: creation or use of institutions or institutionalized processes
- Capacity building: measures to raise awareness and further training of the employees involved and to develop procedures and methods
- Technical data protection: Provision of technical infrastructure for enhanced data protection
- Regulatory approach: highly regulatory measures of the administration / legislator

The different type of measures are also divided into

- short-term,
- medium-term and
- long-term measures

We also divide the data into a traffic light system as follows:

	**Red** = Data record must not be opened in any case **Orange** = records must first pass through a check and may open if necessary, taking into account certain protective measures **Green** = Data set can be opened as raw data (not anonymised)

Remarks for the toolkit:

The approaches presented below are based on very different levels. Some can be implemented very quickly and low-threshold, while others serve as a suggestions for the further development of the subject. The proposals are not isolated, but as elements of one building kit. In principle, they only develop their full effect when combined.

[24] Datensparsamkeit, to reduce data collection, especially from personal data, in the beginning. Datensparsamkeit is a German word that's difficult to translate properly into English. It's an attitude to how we capture and store data, saying that we should only handle data that we really need. https://martinfowler.com/bliki/Datensparsamkeit.html

In general, the more intensively external experts are involved (from the open data as well as the data protection community) and transparency on processes and proceedings is created, the greater the acceptance on all sides of a widespread publication of data from the public sector and business sector[25] - even if something should go wrong.

1. *Before publication:* Decision whether data can or should be open at all

Possible risks that may arise in this decision:

- Data that may or should not be published is incorrectly published (or as a result of insufficient examination).
- The principle of weighing is insufficiently applied and the right to privacy is not sufficiently taken into account.
- The data protection argument is abused in order not to open relevant data sets with high social value.

Examples which illustrate these risks:

- The City of Washington, D.C. Published voter data including name, address and political preferences.[26]
- For a hackathon, the organizers provided mobile phone data (call detail records) on an online platform for download.[27]
- Governments published data on recipients of certain social benefits with names and addresses (Background discussion with NGO representatives from Poland).
- In Bhutan, the data of applicants for public authorities including contact data are provided as open data on the government platform.

Process proposal:
Data providers are informed via materials from a toolbox about the risks of data protection and measures. For the evaluation of the data, the employee is provided with a checklist as a decision aid. This forms a procedure model for data opening and provides the evaluation basis for whether data should be opened or not. The data are thus transferred from the data provider (case worker) based on a traffic light system for the purpose of a risk assessment.

The evaluation procedure must in regular intervals be checked for up-to-dateness and correctness. Red records are not published, green ones are published according to the Open-Definition[28]. Orange records are going through a kind of weakened privacy impact assessment. For classification in safe (green), not too publish (red) and to assess (orange)

[25] The toolkits are based on the work of the SNV and City of San Francisco and were also checked and discussed with the people working on publishing open data at ODINE startups, Deutsche Bahn and members of the Bitkom Open Data Working Group on Bitkom Open Data Taskforce meetings on 15. February 2017 and 27. April 2017.
[26] http://fusion.net/story/314062/washington-dc-board-of-elections-publishes-addresses
[27] https://responsibledata.io/reflection-stories/open-data-hackathon
[28] http://opendefinition.org/od/2.1/en/

the four-eyes principle has to be followed, meaning at least two persons from the respective authority have to come to the same conclusion. If this is not so, the case has to be escalated through a third instance. In addition to this third instance, a data privacy officer (if available) can also be used directly.

For data sets which have a high social relevance or which are demanded intensively, an external advisory body, representatives of the relevant authority, the open data community and data protection experts find a decision to publish.[29]

1.1 Toolbox for open data publishing

Toolbox for open data publishing including, for example, approach, model, notices, contact information etc.

Feasability:	Short-term
Instrument:	Evaluation Assistance
Opportunity / Plus:	• When vividly reprocessed can be good help for administrative staff to identify adequate records • If provided openly or even elaborated collaboratively, this can create trust on the part of the community
Risk / Minus:	• If incomplete, further challenges for data protection will not be considered • Permanent revision is urgently required
Example:	• See „Open Data Release Toolkit“ of the City San Francisco *(highly recommended)* • See „Handreichung Datenschutz“ *(German language)* • See KDZ for the City of Vienna *(German language)*

1.2 Checklists for the assessment of the general data protection risk

Checklists for the assessment of the general data protection risk of the opened up data sets. Ideally this is part of 1.1.

Feasability:	Short-term

[29] For opengov data

Instrument:	Evaluation Assistance
Opportunity / Plus:	• Reduces the uncertainties of the administrative staff • Good material as suggestions from abroad available
Risk / Minus:	• Assessment criteria and standards develop and change over time • The lists must therefore be checked (not often, but regularly) for timeliness and appropriateness • Checklists may contribute to the fact that the assessing person automatically proceeded according to schema X.
Example:	• See „Open Data Release Form“ of the City San Francisco *(highly recommended)*

1.3 Traffic light system for categorizing data sets according to potential privacy risk

Traffic light system for categorizing data sets according to potential privacy risk.
"Red data" should generally not be opened at all.
“Orange” data sets must first be checked and can be opened if necessary, taking into account certain protective measures.
"Green" data set can be opened as raw data (not anonymised).

Feasability:	Mid-term
Instrument:	Evaluation Assistance
Opportunity / Plus:	• Simplification measure for data providers • Increased acceptance if assessment catalog for the categorization is disclosed or even collaboratively created with the civil society
Risk / Minus:	• Risks can (and will) change • Must be constantly updated • Could possibly contribute to downplay risks
Example:	• See suggestion from research Zuiderveen Borgesius et al. (2015) • See 6 steps approach „Datatag System“ by Sweeney et al. (2015) • See „Open Data Release Toolkit“ of the City San Francisco • See PSI-Assessment Austria

1.4 Development of tailor-made solutions for specific topics or authorities

Development of tailor-made solutions for specific topics or authorities for the tools 1.1 - 1.3.

Feasability:	Short-term
Instrument:	Evaluation Assistance
Opportunity / Plus:	• Addressed the challenge of privacy risks in individual authorities or units, since some authorities have more sensitive data (for example health care)
Risk / Minus:	• The risks of individual areas could be difficult to predict in advance. Hence, classification / prioritization is probably only possible with difficulty
Example:	• See „Open Data Release Toolkit“ of the City San Francisco

1.5 Use of a simplified privacy impact assessment for orange data

Use of a simplified Privacy Impact Assessment for orange data. This is ideally part of the toolkit under 1.1)

Feasability:	Mid-term
Instrument:	Evaluation Assistance
Opportunity / Plus:	• When vividly prepared, meaningful - and in other areas established - procedure to identify and evaluate difficult data sets • Increased acceptance if made available for usage or even collaboratively created with the civil society
Risk / Minus:	• Possibly deterrent when only available in "lawyer-lingo" • In order to be able to create appropriate instruments, an extensive bottom-up analysis of existing cases (usually also from the other countries) would be needed • Time delay of data publication • Complexity
Example:	• See as bedrock the Privacy Impact Assessments UK

	• See Bieker et al. (2016). A Process for Data Protection Impact Assessment Under the European General Data Protection Regulation, page 27

1.6 Checking of records to be published according to the four-eyes principle

Checking of records to be published according to the four-eyes principle, being part of 1.1 - 1.4. In case of disagreement a consultation with an independent body is recommended. With orange data at least the data protection officer of the authority has to be involved.

Feasability:	Mid-term
Instrument:	Institutionalization
Opportunity / Plus:	• Avoiding mistakes • (if possible) competences mix
Risk / Minus:	None
Example:	• Is standard procedure and good practice

1.7 Central open data office for all authorities

Creating a Central Open Data Office for all authorities, which assists in the case of uncertainties in the competent authorities (e.g. in the case of disagreement with respect to orange records).
Useful for governments but also the private sector[30].

Feasability:	Mid-term
Instrument:	Institutionalization
Opportunity / Plus:	• A clear contact for data providers • Helpful, especially when technical and data protection expertise is available at the central office

[30] Was very valuable for the German Railway DB

Risk / Minus:	• Binds (still to be created) resources
Example:	• Mindbox at Deutsche Bahn • Currently provided in the framework of the Open Data Law in Germany)

1.8 External Advisory Board

External Advisory Board (ethics committees) which decides on data with great societal importance but also higher data protection risks (e.g. data on refugees, health, smart cities).

Feasability:	Mid-term
Instrument:	Institutionalization
Opportunity / Plus:	• Increased acceptance in society • Strengthens bond with the community • Good for connecting Open Government Partnership activities • Helpful to depict changing minds,the publication of records which is considered to be particularly important at a certain time (such as in the US the publication of police statistics to disclose discrimination) or the classification of previously unproblematic data into risky data (such as with records on refugee shelters)
Risk / Minus:	• Acceptance with administrative staff may not be adequately ensured • Legitimation can be doubted from the outside
Example:	• See as an inspiration the open-government discussion platform The UK Open Government Network. • See the "data sharing consultation process" of the British government • See „Create sector transparency panel" by O'Hara (2011)

2. When publishing: Data protection measures in the course of the publication of data

Possible risks that may arise in this decision:

- Bad or inadequate anonymization
- Possibility of re-identification despite anonymisation

Examples which illustrate these risks:

- After the release of anonymous taxi data in New York, hackers were able to identify the salary of taxi drivers, the movement patterns of celebrities as well as the places of residence of individual passengers.[31]
- The public transport company Transport for London published data on the use of public bicycles, which could be deanonymized and allowed the creation of movement patterns of individual cyclists.[32]
- On an Australian Open Data portal - anonymized data - on prescription was published. The University Melbourne showed that linking these data with other data sets allowed conclusions to be drawn on individual doctors.
- The UK government provided sensitive health data for its citizens on the care.data platform, without sufficient information or to ask for their consent at the beginning. The data access was only possible with registration and the data was pseudonymized, but the privacy of the citizens was de facto only marginally protected, especially as it remained intransparent, who obtained access to the data. The resistance in the population was great and led to the discontinuation of the project.

Process proposal:

In the proposed traffic light system records classified as orange may be published after an anonymisation. To achieve a high degree of anonymisation quick and with high-quality technical instruments must be used. In addition appropriate training for the employees must be offered. Used anonymisation procedures must be evaluated in advance by experts. Those anonymisation procedures will also be used when publishing the data as metadata documents. It is necessary to examine the extent to which privacy-by-design solutions can be integrated into data platforms. In general, the development and implementation of user-friendly technical measures for high-quality anonymisation should be promoted.

2.1 Guidelines for the aggregation and the anonymisation of data

Feasibility:	Short-term
Instrument:	Capacity building
Opportunity / Plus:	● Can provide orientation and work relief, if vividly prepared ● Capacity building for data protection
Risk / Minus:	● The benefit from data is usually inversely proportional to the aggregation level ● Can not rule out errors ● Good anonymization is difficult, therefore questionable whether

[31] https://research.neustar.biz/2014/09/15/riding-with-the-starspassenger-privacy-in-the-nyc-taxicab-dataset
[32] ttp://qz.com/199209/londons-bike-share-program-unwittinglyrevealed-its-cyclists-movements-for-the-world-to-see

	guidelines are enough
Example:	• See the "Code of Practice" for anonymizing by the UK Data Protection Supervisor.

2.2 Anonymisation trainings for data providers

Via courses, online tools, blended learning offers

Feasibility:	Short-term
Instrument:	Capacity building
Opportunity / Plus:	• Increases required competences for technical data protection
Risk / Minus:	• Binds resources and takes time
Example:	• See, for example, the materials of the UK Anonymization Network "Anonymisation Decision-Making Framework" or "Case-Studies".

2.3 Use of technical applications that allow high-quality anonymisation

Use of technical applications that allow high-quality anonymisation, like PET (Privacy Enhancing Technology) which allow users to protect the privacy of their personally identifiable information (PII) provided to and handled by services or applications.

Feasibility:	Mid-term
Instrument:	Technical data protection
Opportunity / Plus:	• Higher data protection level • Empowerment of data providers
Risk / Minus:	• Relying on the technology allows errors to be overlooked more quickly, the capacity building must therefore take place in advance within the office in advance • Applications are often not particularly very user-friendly[33]
Example:	• See the software solutions like

[33] A well-known understatement and major roadblock for acceptance.

	○ ARX – Data Anonymization Tool ○ Aircloak ○ Cornell Anonymization Toolbox ○ Privacy Analytics Risk Assessment Tool ○ University of Texas Anonymisation Toolbox ○ µ-ARGUS – Statistics Netherlands

2.4 Registering the type of anonymization as metadata

Feasibility:	Short-term
Instrument:	Institutionalization
Opportunity / Plus:	● Makes error detection and error avoidance easier ● Corresponding technical procedures could automate this ● Makes the risk assessment according to 3.5 and 3.6 easier
Risk / Minus:	● Possibly increases the risk of specific deanonymization

2.5 External review of anonymisation procedures

Involvement of experts through consultation process to assess anonymisation quality; Re-examination at regular intervals.

Feasability:	Mid-term
Instrument:	Institutionalization
Opportunity / Plus:	● Ensures high and up-to-date standards of anonymisation ● Dynamic adaptation to technical developments ● Strengthens acceptance and trust
Risk / Minus:	● Additional step that binds resources and time costs ● Possibility for delaying the opening of further data
Example:	● See the proposal by the Australian scientists who have selected anonymisation procedures of the authorities to be examined by experts.

2.6 SafeAnswer applications for sensitive data

SafeAnswer applications for sensitive data are technical applications that allow queries to records, but no access to raw data.

Feasibility:	Long-term
Instrument:	Technical data protection
Opportunity / Plus:	• Significantly increased data protection • Also allows the use of sensitive data • Promotion and research of such systems could solve long-term data protection problems
Risk / Minus:	• Still in the research stage and relying on the technology allows errors to be overlooked more quickly • Does not correspond to OpenDefinition (no access to raw data)
Example:	• See the application of SafeAnswer technology at OpenPDS. • See the Differential Privacy approach.

2.7 Privacy by design for special applications and data platforms

Privacy by design for special applications and data platforms, for example automated anonymization.

Feasibility:	Long-term
Instrument:	Technical data protection
Opportunity / Plus:	• Higher data protection level • Empowerment of data providers
Risk / Minus:	• Relying on the technology allows errors to be overlooked more quickly • Anonymization already limits the use potential, meaningful and effective anonymisation always depends heavily on later use
Example:	• See approach of a data usage control, as is pursued in the IND²UCE project by Fraunhofer IESE. See also 3.2 (data access control)

3. *After publication:* Control of the use of already opened data

Possible risks that may arise in this decision:

- Due the described limits of anonymisation, data could be associated with other publicly accessible data, whereby a deanonymization could become possible.
- Open data becomes part of the larger "data ecosystem" and can thus be used without the knowledge of those affected by the profile creation of data marketeers. Thus, as well as other data, such data can be used by insurance or credit providers for pricing options.

Examples which illustrate these risks:

- In Minneapolis the data from car license plates were further processed after the release of data dealers. This led to massive public outrage.[34]
- In Seattle, open government data was used by data traders to combine profiles with other data profiles of citizens to sell them, i.e. to advertisers.[35]

Process proposal:

Open-data platforms are proactively informing about the limits of anonymisation procedures and recommended procedures. Sensitive records are made accessible only with restrictive access. Which records are to be classified as sensitive will be decided in consultation with an external advisory body.

To counter against deanonymization, the spread and use deanonymized records are sanctioned in accordance with the statutory practice. In every publishing office, processes are created how to deal with potential critical data sets. (non-publication, restrictive access, improved anonymization, ...). Cases are collected and evaluated to avoid mistakes in the future and to develop an early warning system. On a regular basis the provided data sets are checked for risks of deanonymization, which can done by internal employees or external experts.

3.1 Training about anonymisation procedures and their limits

Training about anonymisation procedures and their limits on open data platforms and other media channels, where applicable including certification of the platform regarding privacy consideration.

Feasibility:	Short-term
Instrument:	Capacity building

[34] http://openscholar.mit.edu/sites/default/files/dept/files/modernopendataprivacy.pdf
[35] http://btlj.org/data/articles2015/vol30/30_3/1899-1966%20Whittington.pdf

Opportunity / Plus:	• Proactive hints to challenges can have confidence building • Can also be used as a reference in case something goes wrong
Risk / Minus:	• Works only enlightening, but does not minimize the immediate risk
Example:	See explanation of the anonymisation procedures Website of UK Police Data. https://data.police.uk/about/#anonymisation

3.2 Restricted access

Provide restricted data access for registered or even certified users, or on special request for example by researchers.

Feasibility:	Mid-term
Instrument:	Regulatory approach
Opportunity / Plus:	• Higher security standards through verifiability of data users • Provides access to relevant, but risk-proof data sets
Risk / Minus:	• Does not correspond to the Open Definition • Additional bureaucratic effort
Example:	• See the San Francisco approach, according to which some records are only made restrictive („Open Data Release Toolkit“ page 19). • See approach for data access by IND²UCE Fraunhofer IESE, see also 2.7 (data usage).

3.3 Sanctioning the distribution and use of deanonymized data

Feasibility:	Long-term
Instrument:	Regulatory approach
Opportunity / Plus:	• Can limit the transmission of deanonymized data in case of high penalties
Risk / Minus:	• International data flow makes national regulation only partially effective

	• Does not correspond to the Open Definition • Depending on effective enforcement
Example:	• See the recommendation of Australian scientists (against the decision of the Australian government) • Data protection laws already restrict the further use of public data insofar as legitimate interests of persons concerned may not be disproportionately affected. • See also the explanations in the statistic law for the prohibition of intentional re-identification, see §21 of the Bundesstatistikgesetz (German Federal Statistics Act) on the prohibition of re-identification.

3.4 Establishing processes to deal with deanonymization

For example instructions for fast removal of data (including reporting).

Feasibility:	Mid-term
Instrument:	Capacity building
Opportunity / Plus:	• Efficient processes are essential to prevent major risks • Reporting supports acceptance • Good error management helps to avoid similar mistakes in the future
Risk / Minus:	• Acts only as damage control, not against the real risks
Example:	• See here the recommendation from the study of the BMI

3.5 Annual general risk assessment

Annual general risk assessment of open data activities on privacy risks.

Feasibility:	Mid-term
Instrument:	Institutionalization
Opportunity / Plus:	• Dynamic adaptability • Necessary collaboration with science promotes community building
Risk / Minus:	• Requires resources • Limits only the damage

Example:	• See the "Open Data Policy" of Seattle

3.6 Testing for potential external re-identification risks

Continuously forced testing for potential external re-identification risks (and as appropriate the confirmation of the inspection of the open data platforms, similar to certification, see 3.1)

Feasibility:	Mid-term
Instrument:	Institutionalization
Opportunity / Plus:	• Good to get in touch with the community and work together on privacy standards • Dynamic adaptation to technical developments is possible
Risk / Minus:	• Limits only the damage • Requires resources • Presupposes trust in and on the part of the community
Example:	• According to established procedures from the IT security: On the basis of defined codes of ethics, so-called "bug bounty programs" are implemented. The pen-tests are provided by external security experts, hackers or researchers. In Seattle a research team has tested data on re-identifiability.

3.7 Buildup of an error catalog

Cases of privacy infringement are notifiable, are collected and analyzed (internationally), in order to develop an early warning system.

Feasibility:	Long-term
Instrument:	Institutionalization
Opportunity / Plus:	• Effective tool in itself and to build up data protection within the authorities • Helps in combination with 2.4 to better understand errors and avoid future mistakes
Risk / Minus:	• Takes time and must be continuously maintained • Has no immediate effect and results
Example:	• See the specification of the Directive 95/46/EC GDPR, which obliges to report data breaches, see „Notification of a personal

	data breach to the supervisory authority“ Article 33, EU-GDPR.

3.8 General restriction of data usage

In terms of use / licenses

Feasibility:	Mid-term
Instrument:	Regulatory approach
Opportunity / Plus:	• Addresses and continues the current data discourse (varies from country to country) • Decreases for example the risk of discriminatory use of data
Risk / Minus:	• International data flow makes national regulation only partially effective • Does not correspond to the Open Definition
Example:	• See here the archive law, according to which the use of personal data is only allowed for research, not for publishing

Conclusion

The potential of open data is enormous. This must be achieved, however in a way which guarantees the trust of the citizens and protects them from privacy violations or other danger in the long run. It is hardly possible to find a better field to show an example of data usage in common-good oriented and responsible manner.

There is no end to it, when through the provision of data, citizens and consumers are getting the target for data misuse or questionable business models.
At a time when politicians are concerned with the data collection on digital platforms, with profile building, election influence by "fake news" and the like, this should not be a separate mention. In fact many people still find it difficult to understand the privacy relevance when it comes to open data.

The aim of this paper and the subsequent collection of instruments, is not to shy away from the horses, figuratively speaking. Rather, we want to encourage a responsible open-data approach at an early stage. Since the current "cure-all" anonymisation is no longer fully reliable, appropriate protective measures should be introduced from the outset. Especially with OpenData you can and should also involve technical instruments.

Appendix

Recommended further tools and readings for that topic:

Open Data Privacy Playbook
https://cyber.harvard.edu/publications/2017/02/opendataprivacyplaybook

Open Data Priorization Toolkit
https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/1151/filebase/cio_document_library/Open%20Data%20Prioritization%20Toolkit%20Summary.html

Primary Authors: Julia Manske, Tobias Knobloch, Michael Peters, Walter Palmetshofer