D4.2 Legal and privacy toolkit v1

Coordinator: Walter Palmetshofer

With contributions from: Ulrich Atz, Yunjia Li, Daniel Dietrich

Reviewer: Telefonica, Sergio Garcia Gomez

Deliverable nature	Report (R)
Dissemination level (Confidentiality)	Public (PU)
Contractual delivery date	31th July 2015
Actual delivery date	31th July 2015
Keywords	Legal and privacy toolkit

Executive summary

The legal and privacy toolkit is a crucial component for the risk management of the ODINE project.

This deliverable and the final version v2 will focus on research and advice on the state-of-the-art experiences and studies on the legal aspects of public sector information and reuse in Europe. It compiles and publishes a toolkit with guidelines, checklists, and best practice recommendations on the legal considerations of public sector information, as well as open data from other sources. This toolkit is building on existing experiences and surveys of data publication and reuse.

Given the nature of this document, the ongoing changes and external tools, this deliverable will also be available at http://opendataincubator.eu/policy-tool-kit in a more interactive format and will be constantly updated to reflect the demands of the features for this toolkit.

Executive summary
1. Introduction
    1.1 How to use this guide
    1.2 Online version
2. Checklist – Top four things to consider when publishing open data
3. Open data and privacy – managing the risk of publishing data relating to individuals
    3.1 Making data open while considering for privacy
        The privacy risks of open data
    3.2 Reducing the risks of personal identification in open data
        Assessing the risks
        The likelihood for identification breach and potential impact of it
        Reducing privacy risks through de-identification
        Removing identifiers
        Pseudonymisation
        Reducing the precision of the data
        Aggregation
    3.3 Privacy tools list for de-identification
4. Open data, privacy and European Law
5. After publishing great open data
    5.1 For organisations that already publish open data: Map your pathway to open data success
6. Re-use checklist
7. References
8. Appendix
    8.1 Organisations providing advice on anonymisation, data protection, privacy
    8.2 List of European anonymisation events in 2015

1. Introduction

This guide gives a first outline of what to consider when data is published as open data. It sets out what has already been researched in the first months of the project, and provides an plan for the legal and privacy toolkit v2.

The aim for the first version is “What are the critical things to consider when opening up data?” with the focus on privacy of the data based on the demand by current startup applicants of the ODINE program.

1.1 How to use this guide

The Legal and privacy toolkit’s aim is to provide practical advice for all consortium partners of the ODINE project, funded projects and generally interested persons on how to release data. It also covers legal aspects which are crucial for scenarios in which open and enterprise data is used in combination, and what has to be considered for this process and to further push forward open data in Europe.

For that we provide in chapter 2 “Top four things to consider when publishing open data”, a quick overview for the basic things you should consider in the beginning of the process.

After that we provide more detailed input in the further sections.

Disclaimer: The content of this guide does not constitute legal advice. If in doubt, you should always contact a lawyer. For the second version of this toolkit we will get legal support.

1.2 Online version

The latest updates and further material and links will be available at
http://opendataincubator.eu/policy-tool-kit

For questions and suggestion please contact us at policytoolkit@opendataincubator.eu.

2. Checklist – Top four things to consider when publishing open data

1. **Licence**
   - Guide to licensing[1]: At its simplest, open data requires just two things: data and openness. There are lots of aspects to openness, but at its most fundamental, the key is **how the data is licensed**. *Data that doesn't explicitly have an open licence is not open data.*
   - List of licences[2]: This section of the Open Knowledge website lists licenses that are conformant with the principles laid out in the Open Definition.
   - Choose a license for your data: http://choosealicense.com/. As the default **licence for data** we recommend Creative Commons 4.0 or Creative Commons Zero. See details at Legal code[3] understandable for humans.
   - Contribute to the discussion on licensing here[4].
   - Encourage easy re-use of data with the provided licence.

2. **Managing the risk of publishing data relating to individuals**
   - Understanding whether you have data relating to individuals
     - A quick definition of personal data by the EU in section 4
     - A quick intro to 'what is an identifier' in section 3.2
   - Reducing the risk of publishing any personal data by anonymising the data
     - A quick checklist in section 3.2
     - A deeper Anonymisation Decision-making Framework and free online course[5] by the UK Anonymisation Network
     - Free anonymisation workshops coming up in Europe: see section 7.2

3. **Be aware of the legal framework in the EU and your country**
   - EU legislation
   - Implementation on country level
   - Voluntary recommended approaches, e.g.
     - https://responsibledata.io/
     - http://wiki.okfn.org/Personal_Data_and_Privacy

4. **After publishing great open data**
   - Make sure the public knows, see this steps

[1] http://theodi.org/guides/publishers-guide-open-data-licensing
[2] http://opendefinition.org/licenses/
[3] https://creativecommons.org/licenses/by/4.0/legalcode
[4] https://github.com/theodi/open-data-licensing
[5] http://theodi.github.io/ukan-course/

3. Open data and privacy – managing the risk of publishing data relating to individuals

As shown in the current debates[6] privacy and their legal issues and openness are not opposing forces, in fact they are different sides of the same coin and equally important.

Open data advocates often suggest that openness should be the default and that we could and should freely access, use, modify, and share for any purpose. Privacy advocates are on the other side concerned that this openness can lead to a situation where personal information is shared with everyone.[7]

A way could be to see openness and privacy as complementary forces, with privacy as a governing framework to control access to, collection and usage of information basically privacy laws enabling knowledge and control of data about citizens and their surroundings.[8]
On the one side in Europe Swedes have access to tax records, whereas on the other Germans are fighting against such bulk data collection, due to cultural and historical differences. How does all this add up into a set of more or less coherent single European digital market norms so that the European citizens know what they can expect on the legal side? It's understandable why advocates of open data and privacy have different points of view.

But one might be also consider the fact that you want privacy for the same reason you want openness. Because you want to know whether the information held by someone (business, government, ..) on a given problem, or even on you, is verifiable and has benefits.[9] Like 'open by default', privacy is a principle that cuts across all forms of data release. It is fundamentally the same thing. Privacy permits us to share selectively, and grant people access but with limitations.[10]

Therefore we need to include the arguments by privacy advocates in those open data conversations. If we shift our thinking on open data and privacy from one of competing interests to one of a single inextricably linked, albeit complex, issue then we can find a path that enables us to cut a way through the jungle.[11]

We need to agree whether and how open data can include personal information. And we need to stop making a dichotomous distinction between freedom of information laws and data

[6] http://www.freedominfo.org/2015/06/cooperation-urged-by-foi-open-data-privacy-camps/
[7] http://techcrunch.com/2015/06/10/in-the-information-debate-openness-and-privacy-are-the-same-thing/
[8] http://www.freedominfo.org/2015/06/a-2015-online-discussion-openness-and-privacy/
[9] http://buzzmachine.com/2010/09/26/the-benefits-of-publicness/
[10] http://techcrunch.com/2015/06/10/in-the-information-debate-openness-and-privacy-are-the-same-thing/
[11] http://www.freedominfo.org/2015/06/cooperation-urged-by-foi-open-data-privacy-camps/

protection; between privacy and open data policies. We need one single policy framework that controls as well as encourages the use of ‘open’ data.

This delivery and its next version will not remove the tensions between these two opposite points of view, but it might allow us to work toward common objectives and explanations that further each perspective’s work rather than leave us in this standoff.

3.1 Making data open while considering for privacy

In making data open it must also maintain high standards of privacy in the data it releases. The definition of open data means non-personal data. To be crystal clear personal information of private citizens should not be released through open data.

Sometimes there may not be a clear cut distinction between non-personal and personal information and may include de-identified personal information. De-identification of personal information is the removal of obscure personal identifiers and personal information so that identification of individuals, that are the subject of the information, is no longer possible.
In the next sections we provide a quick guide on what has to be considered for that and where we are referencing to the great Privacy and Open Data Guideline guide by the Privacy Committee of South Australia for more details.[12]

The privacy risks of open data

While there are significant economic, democratic and social benefits to the release of government data, it can pose risks to the privacy of personal information. The primary risk to privacy during the release of data is the identification of individuals.
That is releasing personal information or data that can be made into personal information through easily linking with other information.

The violation of an individual privacy leading to identification of an individual person can be significant including humiliation, financial or employment-status impact, depending on the type of data released and the extent of any identification of individuals. This can happen either as spontaneous recognition, which is made without any special effort due to rare characteristics or as deliberate attempt of combining various characteristics and datasets.

[12] http://www.archives.sa.gov.au/sites/default/files/20150121%20Privacy%20and%20Open%20Data%20Guideline%20Final%20V1.1_Copy.pdf

3.2 Reducing the risks of personal identification in open data

Assessing the risks

Assessing the risks of identification of individuals in the release of open data is one of the necessary steps to mitigate those risks to acceptable levels. Following points should be considered:

- Determining of any specific unique identifier like name, date of birth
- Cross-referencing to determine unique combinations like age, gender, ZIPcode, …
- Acquiring knowledge of other publicly available datasets and information that could be used for list matching.

The level of privacy risk will be dependent on the likelihood that identification could occur from the release of the data and the consequences of such a release. The level of risk will determine what steps the agency takes to mitigate the privacy risks.

The likelihood for identification breach and potential impact of it

The likelihood of identification depends on the provided data like name, date of birth or unique identifiers like customer numbers.
Even with such variables missing other factors should to be considered:

- Motivation to attempt identification
- Level of details (the more detail the more likely identification becomes)
- Presence of rare characteristics
- Presence of other information (the dataset itself does not include any data that can identify an individual, but it may include enough variables that can be matched with other information)

Always keep in mind what the potential breach of the privacy could mean for the individual.

Reducing privacy risks through de-identification

Several techniques can be applied to properly de-identify the dataset and reduce any risks of identification of an individual.

Removing identifiers

First step of de-identification is to remove clear identifying variables from

the data (name, date of birth or address).

Removing the identifiers from

Customer	Customerid	Address	Items	Postcode	Annual kilometers	Age
John Doe	23	Street 1	Bike	12345	7500	24

results in

Customerid	Items	Postcode	Annual kilometer	Age
23	Bike	12345	7500	24

While some identifiers are stripped, it retains a relatively high potential for re-identification: the data still exists on an individual level and other, potentially identifying, information has been retained. For example, some ZIPcodes have very small populations and combining this data with other publicly available information, can make re-identification a relatively easy task. While it may be tempting for agencies to strip out all potentially identifying information, doing so could render the data meaningless. The fact that somewhere in Germany there is a railroad customer with the age 24 traveling 7500 km by railroad a year may have limited potential use.

Pseudonymisation

Another method of de-identification is ‘pseudonymisation’ which involves consistently replacing recognisable identifiers with artificially generated identifiers, such as a coded reference or pseudonym. In our example John Doe would be assigned a randomly selected number

Pseudo#	Address	Items	Postcode	Annual kilometers	Age
pseudo123	Street 1	Bike	12345	7500	24

This pseudonymisation allows for different information about an individual, often in different datasets to be correlated without the consequence of direct identification of the individual. For example, the information above could be correlated with:

Pseudo#	Year	Month	Restaurant-Waggon	Food	Age
pseudo123	2015	July	yes	yes	24

Be aware, pseudonymisation also has a relatively high potential for re-identification, as the data exists on an individual level with other potentially identifying information being retained. Also, because pseudonymisation is generally used when an individual is tracked over more than one dataset, if re-identification does occur more personal information will be revealed concerning the individual.

Reducing the precision of the data

Rendering personally identifiable information less precise can reduce the possibility of reidentification. Dates of birth or ages can be replaced by age groups.

Pseudo#	Year	Month	Restaurant-Waggon	Food	Age
pseudo123	2015	July	yes	yes	20-30

Related techniques include suppression of cells with low values or conducting statistical analysis to determine whether particular values can be correlated to individuals. In such cases it may be necessary to apply the frequency rule by setting a threshold for the minimum number of units contributing to any cell. Common threshold values are 3, 5 and 10.
For example, applying a threshold value of 3 to the following table the cell indicating the number of driving instructors at ages 35-40 has a value less than 3 may be suppressed or aggregated into a bigger range.

Age	ZIPcode	train-riders	annual kilometer
20-30	12345	21	<1000
31-40	23456	12	1001-4999
41-50	34567	3	>5000

Introducing random values or ‘adding noise’ is more advanced and may also include altering the underlying data in a small way so that original values cannot be known
with certainty but the aggregate results are unaffected.

Aggregation

Individual data can be combined to provide information about groups or populations. The larger the group and the less specific the data is about them, the less potential there will be for identifying an individual within the group. In our example for aggregating the ZIPcodes on state level.

For further details for this the the sections above we are referencing to the great Privacy and Open Data Guideline guide by the Privacy Committee of South Australia.

3.3 Privacy tools list for de-identification

The following list of tools and software packages are an example for helping de-identifying datasets. These tools provide an automated method of applying a particular de-identification method and may assist an agency to determine with more precision the success of the de-identification method applied and the privacy risk of public release of the dataset.

Cornell Anonymization Toolbox
http://sourceforge.net/projects/anony-toolkit/

Privacy Analytics Risk Assessment Tool
http://www.privacy-analytics.com/

University of Texas Anonymisation Toolbox
http://cs.utdallas.edu/dspl/cgi-bin/toolbox/

μ-ARGUS – Statistics Netherlands
http://neon.vb.cbs.nl/casc/mu.htm

This list will be updated on http://opendataincubator.eu/policy-tool-kit.

4. Open data, privacy and European Law

This section of the delivery aims to assist understanding and addressing the risks to privacy when considering the public release of datasets and has been developed to ensure compliance with the European law. A more detailed version will follow with version 2.

Privacy and data protection are fundamental rights in the EU.
Data protection is a fundamental right, protected by European law and enshrined in Article 8 of the Charter of Fundamental Rights of the European Union.[13]

"Under EU law as well as under CoE law, 'personal data' are defined as information relating to an identified or identifiable natural person, that is, information about a person whose identity is either manifestly clear or can at least be established by obtaining additional information. "
(Data Protection Directive, Art. 2 (a); Convention 108, Art. 2 (a).)[14]

The first EU Data Protection Directive is from 1995. Under this directive, any data "by which an individual can be identified" was the sole responsibility of the data controller, i.e. the owner of this data.
But a newer and stronger regulation is currently on the way, being developed to take into account the vast technology changes since then.
The current plan by the EU is to finalise the regulation in 2015 and implement it by 2017 (similar to the end ODINE project and the version 2 of this delivery). As with any regulation, the current draft could change.

Under the new proposed regulations any business or individual that processes this data will also be held responsible for its protection, including third parties such as cloud providers.[15] To put it simply, anyone who access your data, wherever they are based, is responsible in the case of a data breach. The implications of this are pretty wide, for example third parties will need to be very attentive when it comes to securing the data of others, and data owners will want to thoroughly vet their partners.[16] More specifically, the rules for data protection in the EU institutions - as well as the duties of the European Data Protection Supervisor (EDPS) - are set out in Regulation (EC) No 45/2001. The EDPS is a relatively new but increasingly influential independent supervisory authority with responsibility for monitoring the processing of personal data by the EU institutions and bodies, advising on policies and legislation that affect privacy and cooperating with similar authorities to ensure consistent data protection.

[13] http://fra.europa.eu/en/theme/information-society-privacy-and-data-protection
[14] http://fra.europa.eu/sites/default/files/fra-2014-handbook-data-protection-law_en.pdf
[15] http://www.strategic-risk-global.com/is-your-business-ready-for-the-data-protection-regulation/1408088.article
[16] http://www.cio.de/a/10-things-you-need-to-know-about-the-new-eu-data-protection-regulation,3108536

5. After publishing great open data

Certify your open data: Show that it's easy to find, use and share and describe the

- formats
- metadata
- contact information
- benefit/impact considerations
- business case around open data

This is important to help re-users to understand what it does, how to interpret it correctly, and whether it's appropriate for them to use it.

Further reading: http://theodi.org/guides/what-open-data

5.1 For organisations that already publish open data: Map your pathway to open data success

The Open Data Maturity Model is a way to assess how well an organisation publishes and consumes open data, and identifies actions for improvement.

http://pathway.theodi.org/

6. Re-use checklist

What to consider when you are re-using data:

1. Make sure you check the **license** and if there are any limitations which could reduce your planned usage of the data.
2. Make sure you do the required **attribution** and citation. Also it's always nice to give that, even if it's not needed.
3. Make sure that the data **doesn't** allow to identify individuals if you combine various data sources. See section 3.2

For a detailed guide see here Reuser's Guide to Open Data Licensing.[17]

A checklist for combining re-using various datasources will come with the next update.

[17] https://theodi.org/guides/reusers-guide-open-data-licensing

7. References

De-Identification
http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

Handbook on European data protection law
http://fra.europa.eu/sites/default/files/fra-2014-handbook-data-protection-law-2nd-ed_en.pdf

Guidelines on recommended standard licences, datasets and charging for the reuse of documents (2014/C 240/01)
http://ec.europa.eu/newsroom/dae/document.cfm?action=display&doc_id=6421

OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data
https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

Open Data Handbook
http://opendatahandbook.org/guide/en/how-to-open-up-data/

Open Data World Bank Blog
http://blogs.worldbank.org/opendata/how-can-the-open-government-data-toolkit-help-you

Opening a new Chapter for Data Protection
https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/EDPS/PressNews/Press/2015/EDPS-2015-06-EDPS_GDPR_EN.pdf

Open Definiton
http://opendefinition.org/ and http://opendefinition.org/guide/data/

Protection of personal data
http://ec.europa.eu/justice/data-protection/

Reuser's Guide to Open Data Licensing
https://theodi.org/guides/reusers-guide-open-data-licensing

Privacy debate links
http://www.freedominfo.org/2015/06/cooperation-urged-by-foi-open-data-privacy-camps/
http://www.freedominfo.org/2015/06/a-2015-online-discussion-openness-and-privacy/
http://techcrunch.com/2015/06/10/in-the-information-debate-openness-and-privacy-are-the-same-thing/
http://buzzmachine.com/2010/09/26/the-benefits-of-publicness/

8. Appendix

8.1 Organisations providing advice on anonymisation, data protection, privacy

- UKANON http://ukanon.net/external-resources/
- HARVARD Privacy Tools http://privacytools.seas.harvard.edu/
- EFF http://eff.org
- EDRI http://edri.org
- Access now http://accessnow.org
- Chaos Computer Club http://ccc.de/
- Cornell Anonymization Toolbox http://sourceforge.net/projects/anony-toolkit/
- Privacy Analytics Risk Assessment Tool http://www.privacy-analytics.com/
- University of Texas Anonymisation Toolbox http://cs.utdallas.edu/dspl/cgi-bin/toolbox/
- µ-ARGUS – Statistics Netherlands http://neon.vb.cbs.nl/casc/mu.htm

8.2 List of European anonymisation events in 2015

12 Aug 2015 Free anonymisation workshop in the UK http://ukanon.net/events/
01-02 September 2015 Workshop on Anonymisation of personal FOT data, in Gothenburg http://fot-net.eu/event/workshop-on-anonymisation-of-personal-fot-data/
25 Sep 2015 Free anonymisation workshop in the UK http://ukanon.net/events/
27 Oct 2015 Free anonymisation workshop in the UK http://ukanon.net/events/
11 Nov 2015 Free anonymisation workshop in the UK http://ukanon.net/events/
10 Dec 2015 Free anonymisation workshop in the UK http://ukanon.net/events/

For an updated list see http://opendataincubator.eu/policy-tool-kit.