International Journal of Spatial Data Infrastructures Research, 2012, Vol.7, 196-206

Brave New Open Data World?*

Stefan Kulk[1], Bastiaan van Loenen[2]

[1] Utrecht University; s.kulk@uu.nl

[2] Delft University of Technology; b.vanloenen@tudelft.nl

Abstract

There is a growing tendency to release all sorts of public data on the Internet. The greater availability of interoperable public data catalyses secondary use of such data, which leads to growth of information industries and better government transparency. Open data policies may, nevertheless, be in conflict with the individual's right to information privacy as protected by the Data Protection Directive. This directive sets rules to the processing of personal data in the European Union. Technological developments and the increasing amount of publicly available data are, however, blurring the lines between non-personal and personal data. Open data may not seem to be personal data on first glance especially when it is anonymised or aggregated. However, it may become personal data by combining it with other publicly available data or when it is de-anonymised. In this article, we argue that these developments extend the reach of European Union privacy regulation to open data, which may obstruct the implementation of open data policies in the European Union.

Keywords: Open data, privacy legislation, European Union

*This work is licensed under the Creative Commons Attribution-Non commercial Works 3.0 License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/ or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

DOI: 10.2902/1725-0463.2012.07.art10

1. INTRODUCTION: OPEN DATA

There is a growing tendency to release all sorts of public data on the Internet. Regional, national and local governments allow individuals and companies to freely re-use public data for their own purposes, often referred to as 'open data'. Such data could, for instance, concern: crime statistics, energy consumption, addresses in a trade register and issued construction permits. A large majority of these data relates to locations on earth; so-called geographic data.

The European Union advocates the re-use of public data through the Directive on the Re-use of Public Sector Information. European Commissioner Kroes has said that she is a fan of open data (Kroes, 2011), and the European Commission strongly advocates open data (European Commission, 2011). The Commission's hopes are that the greater availability of interoperable public data catalyses secondary use of such data, which leads to growth of information industries and better government transparency. The re-use of public data has great commercial potential. The total value of re-use of public sector information in Europe is estimated to vary from €27 billion (Dekkers et al., 2006), €68 billion (Pira International, 2000) up to € 140 billion (Vickery, 2011).

Open data policies may be in conflict with the individual's right to information privacy as protected by the Data Protection Directive, that sets rules to the processing of personal data in the European Union. Technological developments and the increasing amount of publicly available data are, however, blurring the lines between non-personal and personal data. These developments extend the reach of EU privacy regulation to open data and could in effect obstruct the implementation of open data policies in the EU. This article discusses the impact privacy legislation in Europe may have on open government data initiatives and policies.

2. OPEN GOVERNMENT DATA

There exist many definitions of open data or open government data. Here we follow the definitions of the open government data working group (see http://opengovernmentdata.org). This group of 30 experts agreed on the following 9 principles. Government data shall be considered open if the data are made public in a way that complies with the principles below:

1. Data Must Be Complete

All public data are made available. Data are electronically stored information or recordings, including but not limited to documents, databases, transcripts, and audio/visual recordings. Public data are data that are not subject to valid privacy, security or privilege limitations, as governed by other statutes.

2. Data Must Be Primary

Data are published as collected at the source, with the finest possible level of granularity, not in aggregate or modified forms.

3. Data Must Be Timely

Data are made available as quickly as necessary to preserve the value of the data.

4. Data Must Be Accessible

Data are available to the widest range of users for the widest range of purposes.

5. Data Must Be Machine processable

Data are reasonably structured to allow automated processing of it.

6. Access Must Be Non-Discriminatory

Data are available to anyone, with no requirement of registration.

7. Data Formats Must Be Non-Proprietary

Data are available in a format over which no entity has exclusive control.

8. Data Must Be License-free

Data are not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed as governed by other statutes.

9. Compliance must be reviewable.

A contact person must be designated to respond to people trying to use the data and to respond to complaints about violations of the principles. An administrative or judicial court must have the jurisdiction to review whether the agency has applied these principles appropriately.

In addition to the principles of the open government data working group we add a 10th principle from the Open Knowledge Foundation (see http://opendefinition.org/okd):

10. The work shall be available as a whole and at no more than a reasonable reproduction cost, preferably downloading via the Internet without charge.

3. DATA PROTECTION IN THE EUROPEAN UNION

Article 8 of the Charter of Fundamental Rights of the European Union gives everyone the right “to the protection of personal data concerning him or her”. The automated processing of personal data is also bound to the Council of Europe’ s Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. This fundamental right is further elaborated by the Data Protection Directive (European Union, 1995). The directive applies to the processing of personal data. When personal data are processed, they should be processed fairly, lawfully and for specified, explicit and legitimate purposes.

3.1. The concept of personal data

The Data Protection Directives defines personal data as “information relating to an identified or identifiable natural person”. An identifiable person is “one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity” (Article 2(a) of the Data Protection Directive).

Typical examples of personal data are names, (e-mail/IP) addresses and telephone numbers. Personal data are, however, more than just names and addresses. The Article 29 Working Party, the group of European Data Protection Agencies, emphasizes the purpose or result that information has, in order to determine whether information is personal data: “data relates to an individual if it refers to the identity, characteristics or behaviour of an individual or if such information is used to determine or influence the way in which that person is treated or evaluated" (Article 29 Working Party, 2005, p. 8).

Personal data includes any sort of statement about a person. Not only objective information is included, but also subjective information. The information does not need to be true. The statement that ‘Titius is a reliable borrower’ is personal data, even if Titius proves to be an unreliable borrower (Article 29 Working Party, 2007, p. 6).

On some occasions information concerning objects can be personal data. Objects usually belong to someone. For instance, the value of a house is information about an object to which on first glance data protection rules do not apply. However, houses are assets of their owners and their value can be used to determine the extent a person’s obligation to pay taxes. In this context, the value of a house becomes personal data to which the Data Protection Directive applies (Article 29 Working Party, 2007, p. 9).

The assessment whether data should be considered personal data depends also on how easy it is to link data to a person (recital 26 of the Data Protection Directive). The stage of technological development is an important measure in this assessment. Data that today is considered not to be personal data may next year very well be personal data. Simply because technological developments have made it possible to identify persons in that data. One example may be the publication on the Internet of a picture including anonymous individuals. Ten years ago, it was almost impossible to uncover the identity of the people in the picture (thus no personal data). Today, image recognition software allows to identify these people with a simple mouse click (thus personal data). Since it is very difficult to effectively remove data from the Internet once it has been put online, one could argue that any data that in the future may be linked to individuals, should be considered and treated as personal data.

The legislative history of the Dutch law implementing the Data Protection Directive can serve as an example. First, in 1998 ownership and value information of real estate was not considered personal data (Kamerstukken 25892 nr. 3). Also in 1998, data at the zip-code level was not considered to be personal data (Kamerstukken 25892 nr. 6). However, in 1999, it was argued that data at the ZIP-code level (6PPC) (see Kamerstukken 25892 nr. 92c) and ownership and value information of real estate should be considered personal data (Kamerstukken 25892, nr. 9).

3.2. Processing personal data requirements

If personal data are processed then the Data Protection Directive provides the requirements that need to be fulfilled.

The Data Protection Directive rules that personal data should be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes (Article 6(1)(b) Data Protection Directive). The processing should further be adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed (Article 6(1)(c) Data Protection Directive).

The controller must also ensure that the processing of personal data is accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified (Article 6(1)(d) Data Protection Directive).

Finally, personal data must be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data are processed (Article (6)(1)(e) EU Privacy Directive).

4. IS OPEN DATA PERSONAL DATA?

Train times, the location of public toilets and the number of car accidents could all be open data. No open data provider is likely to offer names, addresses, social security numbers, or other data that directly or indirectly identifies natural persons as open data. Open data (ideally) is at the most anonymised or aggregated data that cannot be related to individuals. The Open Knowledge Foundation visualizes open data and “private data” as two non-overlapping subsets. Unfortunately, in reality this distinction is not so easy to draw. Even when data has been anonymised or aggregated, data analysis techniques now allow us to re-identify individuals in such data (Ohm, 2010). This is not just a theoretical problem.

When AOL released 650,000 pseudonymised search queries, two New York Times journalists showed that some search queries could be related to a individual (Barbaro et al., 2006). The journalists had uncovered the identity of 62-year old widow Thelma Arnold living in Lilburn, Ga., USA. She searched for topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything”. The widow’s identity could be revealed because she had searched for people with the last name Arnold, and issues relating to her hometown.

Another example is the Netflix-case. Netflix offered anonymised data for a contest to improve its movie recommendations. Researchers showed that these data could be linked to certain Netflix subscribers (Narayanan et al., 2008). The researchers demonstrated that someone who knows only a few things about an individual Netflix subscriber can easily identify this subscriber’s record in the dataset. Using data from the Internet Movie Database as a source of background knowledge, Narayan and Shmatikov successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Another research project demonstrated that it is relatively easy to link open data to individuals. Students uncovered names and addresses of councillors, and names, posts and salaries of senior public servants by combining data from the British open data portal with other already available public data (Simpson, 2011).

A study of the 1990 US census data showed that 87% of the US population could be identified based on only gender, 5-digit ZIP code and date of birth (Sweeney 2000). A similar study was conducted on the 2000 US census data, which made clear that 63% of the US population could be identified by only those same characteristics (Golle, 2006).

These examples show us that anonymised or aggregrated data that is non-identifiable today, may turn out to be indirectly identifiable tomorrow. Referring to the case of AOL, one could argue that the dataset should have been better

anonymised by, for instance, removing names and other identifying information from the search queries. In other words, identification could have been prevented. However, in the cases of Netflix and the British open data portal, there were no names or other directly identifying data included in the datasets. Combining the datasets with already available information lead to identification of individuals in certain data. This would have been difficult to prevent.

The risk of de-anonymisation or de-aggregation of data will only grow as computing power increases and more data becomes publicly available. Does this mean that anonymised or aggregated open data are or will become personal data? The answer to this question depends on two issues.

Firstly, according to recital 26 of the Data Protection Directive, all the means likely reasonably to be used to identify a person should be taken into account to determine whether a person is identifiable. The Data Protection Directive does not apply to data that is rendered anonymous in such a way that the person is no longer identifiable. According to the Article 29 Working Party, the assessment of whether data allows identification of an individual, and whether the information can be considered as anonymous, depends on the circumstances of the case. A case-by-case analysis should be carried out taking into account the means that are likely to be used for identification (Article 29 Working Party, 2007, p. 21).

Secondly, in order for open data to be personal data, the data in its de-anonymysed or de-aggregated form should relate to an individual. For instance, de-aggregated data that relates to the number of house sparrows in a certain area will most likely not relate to individuals and will not be personal data. On the contrary, if the de-anonymised or de-aggregated data relates to the value of a specific real estate or the use of electricity on household level, the data should be considered personal data.

Whether open data are personal data depends on the circumstances of a particular case. In the case of the Netflix-prize, researchers managed to uncover sensitive information such as political preferences. Such information is personal data in the sense of Article 8(1) of the Data Protection Directive, which explicitly prohibits the processing of special categories of personal data such as “political opinions”. Because the information was uncovered by two specialised computer scientists and the directive applies to information that can be identified by “all the means likely reasonably to be used”, it is unsure whether, at this moment, the information would really be personal data in the sense of the directive. In contrast, the British open data can be qualified as personal data for it were students who uncovered names, addresses and salaries from the British open data.

5. OPEN DATA ARE PERSONAL DATA. NOW WHAT?

Processing of personal data is not by definition unlawful. The Data Protection Directive requires that if personal data are processed, it should be done fairly, lawfully and for specified, explicit and legitimate purposes (Article 6 of the Data Protection Directive). The purposes for which the data are processed must be explicit and legitimate and must be determined at the time of collection of the data (Recital 28 of the Data Protection Directive).

Especially the requirement of 'specified' purposes will give rise to problems when open data are personal data. The purpose of open data policies is to allow access and re(use) of data without any or with very few limitations. Unconditional (re)use of data is not specific enough to fulfil the requirement of a specified purpose. This requirement protects data subjects as it requires data processors to limit the purposes of the data processing. It also allows data subjects to gain insight into what is done with data that relates to them. A purpose that is as broad as the open data purpose is in fact a blank cheque and does not allow data subjects to know or control what their data is used for.

The processing of certain types of personal information is in principle prohibited. This is the case when the processing of the information reveals racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and information concerning health or sex life (Article 8(1) Data Protection Directive).

This prohibition may also be a problem for open data-providers that want to offer aggregated data concerning, for instance, political opinions in certain neighbourhoods. They process sensitive personal data, which is prohibited if the data can be de-anonymised or de-aggregated.

6. CONCLUSION

Open data policy catalyses the sharing of geographic data, which brings great economic benefits and could help make government institutions more transparent. However, at the same time open data puts data protection under pressure and will increasingly do so as technology progresses and more data becomes publicly available.

Directly identifying data, such as names, social security numbers and credit card numbers, will not be published in open data initiatives. A data provider may choose to publish aggregated or anonymised data. However, even if data are aggregated and anonymised, it is relatively easy to link these data to individuals, making them personal data in the context of the Data Protection Directive. This directive requires that personal data are processed for specified, explicit and

legitimate purposes and not further processed in a way incompatible with those purposes. However, open data initiatives by definition lack an explicit and specific goal for the data processing. This leads to a simple, but as complex, conclusion: European privacy legislation is a serious barrier for open data initiatives. Recently, the Europan Data Protection Supervisor has recognised the link between open data and privacy and calls for a prohibition on re-identification of data-subjects, among other recommendations (see EDPS, 2012).

7. FUTURE RESEARCH

Technological developments have always challenged the protection of the right to information privacy. Now that converging and ambient technologies become more ingrained in our society, the question whether existing European data protection law is sufficient to deal with new technological developments has gained more prominence. De Hert speaks of normative and technological discrepancies between practice and data protection law (de Hert, 2009). Koops calls data protection laws outdated and unable to protect the citizen's right to privacy (Koops, 2011). He raises, but does not answer, the question whether the future of information privacy should be sought outside the realm of privacy and data protection law itself.

This question can also be raised in the context of open data. Is EU privacy regulation still up to date when it creates a barrier to open data policy that seems impossible to break? Ideally, EU data protection law strikes a balance between the right to information privacy and open data that still allows to reap the benefits of open data policy. This issue should be investigated in greater depth.

ACKNOWLEDGEMENTS

The research presented in this article was supported by the Dutch Innovation program Next Generation Infrastructures.

REFERENCES

Article 29 Working Party (2005). Working Document on Data Protection Issues Related to RFID technology.

Article 29 Working Party (2007). Opinion 4/2007 on the Concept of Personal Data.

Barbaro, M. and T. Zeller (2006). A Face Is Exposed for AOL Searcher No. 4417749. *The New York Times*, at: http://www.nytimes.com/2006/08/09/technology/09aol.html [accessed 14 February 2012].

Dekkers, M., F. Polman, R. te Velde & M. de Vries (2006). *Measuring European Public Sector Information Resource. Final Report of Study on Exploitation of public sector information – benchmarking of EU framework conditions*.

EDPS (2012). Opinion of the European Data Protection Supervisor. Opinion of the European Data Protection Supervisor on the 'Open-Data Package' of the European Commission including a Proposal for a Directive amending Directive 2003/98/EC on re-use of public sector information (PSI), a Communication on Open Data and Commission Decision 2011/833/EU on the reuse of Commission documents, http://www.edps.europa.eu [accessed 25 April 2012].

European Commission (2011). A Digital Agenda for Europe (COM(2010) 245 final/2).

European Union (1995). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (OJ L 281/95).

European Union (2003). Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information (OJ L 345/90).

Golle, P. (2006). "Revisiting the uniqueness of simple demographics in the US population", *WPES '06 Proceedings of the 5th ACM workshop on Privacy in electronic society*.

de Hert, P.J.A. (2009). *Citizens' data and technology: An optimist perspective*, The Hague: Dutch Data Protection Authority.

Koops, B.J. (2011). The Evolution of Privacy Law and Policy in the Netherlands *Journal of Comparative Policy Analysis: Research and Practice*, 2011-2, p. 165.

Kroes, N. (2011). Public data for all, at: http://blogs.ec.europa.eu/neelie-kroes/public-data-for-all---opening-up-europes-public-sector [accessed 14 February 2012].

Narayanan, A. and V. Shmatikov (2008). Robust de-anonymization of large sparse datasets, *Security and Privacy,* 2008-5, p. 111.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization, *UCLA Law Review,* 2010, p. 1701

Pira International (2000). *Commercial exploitation of Europe's public sector information*.

Simpson, A.C. (2011). On privacy and public data: a study of data.gov.uk, *Journal of Privacy and Confidentiality,* 2011-1, p. 4.

Sweeney, L. (2006). “Uniqueness of simple demographics in the U.S. Population”, *LIDAPWP4*, Carnegie Mellon University.

Vickery, G. (2011). Review of recent studies on PSI re-use and related market development. Paris, 41 pages.