Open Government Data Publication Methodology

Jan Kucera
Faculty of Informatics and Statistics
University of Economics Prague

Abstract: Public sector bodies hold a significant amount of data that might be of potential interest to citizens and businesses. However the re-use potential of this data is still largely untapped because the data is not always published in a way that would allow its easy discovery, understanding and re-use. Open Government Data (OGD) initiatives aim at increasing availability of machine-readable data provided under an open license and therefore these initiatives might facilitate re-use of the government data which in turn might lead to increased transparency and economic growth. However the recent studies show that still only a portion of data provided by the public sector bodies is truly open. Public sector bodies face a number of challenges when publishing OGD and they need to address the relevant risks. Therefore there is a need for best practices and methodologies for publication of OGD that would provide the responsible persons with a clear guidance on how the OGD initiatives should be implemented and how the known challenges and risks should be addressed.

Key words: Best Practice, Challenges, Methodology, Open Data, Open Government Data, OGD, Public Administration

1. Introduction

Public sector bodies (hereafter PSBs) hold a significant amount of data that might be of potential interest to citizens and businesses. Data is a resource with a huge potential to re-use but this potential is yet largely untapped. Partly due to the fact that data is often not shared in a way that would allow its easy discovery, understanding and re-use. (G8, 2013)

According to Open Knowledge Foundation (© 2010-2012) Open Data is “data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.” Open Data should be available under a license/terms of use that enable the above mentioned free use, re-use and redistribution of the data by any interested subject and it should be provided in a suitable machine-readable format (Open Knowledge Foundation, © 2010-2012). Therefore publishing the data held by PSBs as Open Government Data (hereafter OGD) might help to exploit its re-use potential (Deloitte, 2012; Ubaldi, 2013).

Publishing and re-use of OGD could result into significant benefits such as increased transparency, improved government services and stimulation of economic growth (Logica Business Consulting, 2012). According to the Open Knowledge Foundation (© 2010-2012) “keep it simple” principle should be followed when opening up data, but for example Janssen & Zuiderwijk (2012) observed that in practice it might be difficult to open up particular datasets because issues such as the confidentiality, data quality or the privacy infringement risks need to be addressed. Besides the privacy infringement risk, there might be other risks associated with the publication of OGD as well, e.g. publication of data against the law or possible misinterpretation of the data (Kučera & Chlapek, 2014). Even though OGD initiatives have been launched in many countries across the globe, according to the World Wide Web Foundation (2015) only over 10% of the 1.290 datasets surveyed in the second edition of the Open Data Barometer study were published under an open license, in bulk and in machine-readable formats. Ubaldi (2013) points out that there are not only technical and legal challenges associated with the OGD initiatives but there are also challenges related to policy, financing, organization and culture.

The abovementioned challenges and risks show that there is a need for best practices and methodologies for publication of OGD that would provide the responsible persons with a clear guidance on how the OGD initiatives should be implemented and how the know challenges and risks should be addressed. If the challenges are not properly tackled it might prevent the expected benefits from being reaped (Ubaldi 2013). The goal of this paper is to discuss the emerging best practices for publication of OGD and to propose the OGD publication methodology.
This paper is structured as follows. Introduction is followed by the description of the research approach. The OGD movement is discussed in the following section. The next section discusses the current best practices for the OGD publication. Open Government Data publication methodology is proposed in the following section. Concluding remarks are presented at the end of this paper.

2. Research Approach

Meaning of the term “methodology” is twofold: the science of methods and a set of methods used in a particular activity (Checkland, 2000). In this paper the latter meaning of the term “methodology” is followed. A method represents “a particular way of doing something” (Oxford Advanced Learner’s Dictionary, 2010, p. 965) or “a defined, repeatable approach to address a particular type of problem” (Open Group, © 1999-2011). According to Checkland (2000) a methodology represents a generic framework when related to method. However when a methodology is used or adapted to solve a particular problem it becomes a method (Checkland, 2000).

The problem domain addressed in this paper is the publication of Open Government Data. OGD publication methodology can be defined as a set of methods, techniques or practices for publication of Open Government Data. i.e. the OGD publication methodology provides recommendations or guidelines how Open Government Data should be published.

Design Science Research Methodology (DSRM) as described by Peffers et al. (2007) was applied to design the OGD publication methodology proposed in this paper (hereafter MePOD-VS methodology¹). DSRM consists of six activities: (1) Identify problem and motivate, (2) Define objectives of a solution, (3) Design and development, (4) Demonstration, (5) Evaluation and (6) Communication. DSRM proposes an iterative research process where the outcomes of the evaluation or the communication activity might trigger new iterations of the process (Peffers et al., 2007). Original version of the MePOD-VS methodology is described in (Kučera, 2015). Slightly updated version of the methodology is presented in this paper that implements some of the findings obtained during the evaluation of the original version of the MePOD-VS methodology. Application of the DSRM to the development of the MePOD-VS methodology as described in (Kučera, 2015) is summarized in table 1.

Tab. 1: DSRM activities in the design of the MePOD-VS methodology, source: author

<table>
<thead>
<tr>
<th>DSRM Activity</th>
<th>MePOD-VS design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify problem and motivate</td>
<td>Literature review, analysis of the results of a project aimed at development of</td>
</tr>
<tr>
<td></td>
<td>the OGD publication plan of one of the Czech PSBs, assessment of the existing</td>
</tr>
<tr>
<td></td>
<td>methodologies.</td>
</tr>
<tr>
<td>Define objectives of a solution</td>
<td>OGD publication methodology content requirements and structural requirements</td>
</tr>
<tr>
<td></td>
<td>specification, update of the requirements according to the results of the</td>
</tr>
<tr>
<td></td>
<td>evaluation.</td>
</tr>
<tr>
<td>Design and development</td>
<td>Design of the MePOD-VS meta-model. Development of the methodology contents</td>
</tr>
<tr>
<td></td>
<td>(principles, roles, processes, inputs and outputs of the processes, relevant</td>
</tr>
<tr>
<td></td>
<td>standards). Description of the methodology implementation process.</td>
</tr>
<tr>
<td>Demonstration</td>
<td>MePOD-VS methodology was utilized in the COMSODE project² where it served as</td>
</tr>
<tr>
<td></td>
<td>an input during the development of the international Methodology for publishing</td>
</tr>
<tr>
<td></td>
<td>datasets as open data (hereafter COMSODE methodology, see (Nečaský et al., 2014)).</td>
</tr>
<tr>
<td>Evaluation</td>
<td>The following methods were used to evaluate the MePOD-VS methodology:</td>
</tr>
<tr>
<td></td>
<td>• Comparison with the approach applied by one of the Czech PSBs during its OGD</td>
</tr>
<tr>
<td></td>
<td>initiative.</td>
</tr>
<tr>
<td></td>
<td>• Comparison with the early results of a project where the COMSODE methodology</td>
</tr>
<tr>
<td></td>
<td>has been applied.</td>
</tr>
<tr>
<td></td>
<td>• Assessment of the methodology design based on interviews with the OGD</td>
</tr>
<tr>
<td></td>
<td>practitioners and researchers.</td>
</tr>
<tr>
<td></td>
<td>• Assessment of the methodology design based on the factors facilitating the</td>
</tr>
<tr>
<td></td>
<td>PSI publication and re-use described in (Kučera et al., 2014).</td>
</tr>
</tbody>
</table>

¹ Same identifier of the methodology as in (Kučera, 2015) is used in this paper.
² http://www.comsode.eu/
<table>
<thead>
<tr>
<th>DSRM Activity</th>
<th>MePOD-VS design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>Assessment of the existing OGD publication methodologies using the content requirements was presented in (Chlapek et al., 2014). Befits and risks of OGD are discussed in (Kučera &amp; Chlapek, 2014). Some of the processes proposed in the MePOD-VS methodology were also discussed during a tutorial (Kučera, 2014a; Kučera, 2014b).</td>
</tr>
</tbody>
</table>

PSBs face a number of challenges when publishing OGD (Janssen et al., 2012; Kučera et al., 2015; Ubaldi, 2013). Problems or issues that the MePOD-VS methodology tries to address were identified during the literature review and some of them were also identified during a project that involved analysis of data sources, identification of the potential datasets for publication and development of the OGD publication plan of a Czech PSB. Topics that should be covered by an OGD publication methodology (content requirements) were derived from the identified problems and issues. A set of suitable elements\(^3\) (e.g. roles, processes, tasks etc.) were selected in order to properly represent the content of the methodology (structural requirements). Using these requirements a meta-model of the MePOD-VS methodology was designed. Recommendations that form the methodology are therefore structured according a common meta-model.

3. Open Government Data

Two main attributes of Open Data have already been introduced – the machine-readability of the data (technical openness) and the open licensing (legal openness), see for example (Open Knowledge Foundation, © 2010-2012). The Open Definition (Open Knowledge, 2014) provides more precise explanation of the “openness” concept that is a part of the Open Data phenomenon. According to the Sunlight Foundation (2010) OGD should not only be machine-readable and openly licensed but it should be also complete, primary, timely, easily accessible, non-discriminating, using commonly owned standards, permanent and available at minimal or no usage costs.

According to Bauer & Kaltenböck (2011) the aim of the OGD initiatives is to open up data produced and commission by the PSBs for re-use by citizens, academia, businesses as well as the PSBs themselves. OGD facilitates the access to data and information held by governments and therefore it is seen as one of enablers of the Open Government movement which aims at establishing “modern cooperation among politicians, public administration, industry and private citizens by enabling more transparency, democracy, participation and collaboration” (Bauer & Kaltenböck, 2011). Besides the Open Government movement, OGD is also related to the freedom of information (FOI) legislation and the European legislation on re-use of the public sector information (PSI), even though it does not build upon any of these legislative domains (Janssen, 2011).

Free access to government information is often seen as one of the basic rights. According to OECD (2005) 28 of the 30 OECD member states had implemented the freedom of information legislation by 2004. Janssen (2011) points out that the increased transparency and accountability are common goals of both the FOI legislation and OGD. While the FOI puts emphasis on obligations of the PSBs to respond to requests for information, the OGD movement stresses the proactive release of data in machine-readable formats under conditions allowing its re-use (Access Info Europe & Open Knowledge Foundation, 2011). However the recent development shows that the FOI advocates put increased emphasis on proactive release of information and that the FOI requests are being used to get an access to the entire databases (Access Info Europe & Open Knowledge Foundation, 2011). While this convergence might demonstrate importance of the OGD initiatives, Janssen (2011) points out to a potential risk in overemphasizing the provision of the data in machine-readable formats because only developers and activists might be able to re-use it. This target group is significantly smaller than the general public. Therefore Janssen (2011) advocates that governments and PSBs should not “lose sight of their existing tasks to inform the public.”

In order to facilitate re-use of the data and information held by the PSBs the European Union adopted the directive 2003/98/EC (PSI Directive) in 2003 (European Union, 2003) which was amended by the directive 2013/37/EU in June 2013 (European Union, 2013). The aim of the PSI Directive is to stimulate European information market and the main target group of the PSI Directive are businesses creating products and services based on the provided information (Janssen, 2011). Fostering innovation and economic growth are the common goals of both the OGD initiatives and the PSI

\(^3\) For example elements of the software development methodologies are discussed in (Cockburn, 2000).
Directive (Janssen, 2011). However the PSI represents a broader concept when compared to OGD because documents that fall within the scope of the PSI Directive represent “any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording)” (European Union, 2003). Nevertheless the PSI re-use is broadly defined and thus in most cases the re-use of OGD should be considered as the PSI re-use (Janssen, 2011). Therefore in situations when publication of machine-readable data is appropriate, publication of such data as OGD might be a way in which the data can be made available in accordance with the requirements of the PSI Directive.

The European Commission supports both the publishers and the re-users of OGD. The Open Data Support project is aimed at the OGD publishers and it provides training, IT advisory and consultancy services as well as data and metadata preparation, transformation and publication services (Loutas, 2013). Open Data startups and SMEs will be supported by the Open Data Incubator for Europe (Open Data Incubator for Europe, © 2015).

OGD is a global phenomenon. The previous paragraphs were dedicated to the European legislation and the support of the European Commission. However the European PSI Directive is not the only legislation relevant to OGD. For instance in the US the Open Data Policy was issued in 2013 (Executive Office of the President, 2013). OGD initiatives in 86 countries were analyzed in the 2nd edition of the Open Data Barometer study (World Wide Web Foundation, 2015). According to this study the United Kingdom and the United States followed by Sweden, France, New Zealand, Netherlands, Canada, Norway, Denmark, Austria and Germany represent the top leading countries in OGD readiness, implementation and impact (World Wide Web Foundation, 2015).

4. Emerging Best Practices for Publishing Open Government Data

As a response to the need to tackle the challenges and problems related to OGD publication and re-use, OGD publication methodologies are being developed. Number of such methodologies has already been proposed. Some of them are being developed by international organizations like United Nations (2013) or World Bank (2014). However there are OGD publication methodologies being developed by governments, business entities and researchers as well. For more detailed discussion of the existing methodologies see for example (Chlapek et al., 2014; Kučera et al., 2015).

Apart from the OGD publication methodologies, recommendations for publication of OGD might also have a form of best practices. According to Gartner (© 2013) a best practice is defined as “a group of tasks that optimizes the efficiency (cost and risk) or effectiveness (service level) of the business discipline or process to which it contributes. It must be implementable, replicable, transferable and adaptable across industries.” Best practice could be seen as a way of doing a particular activity that is recommended to be followed in order to achieve the desired results. Best practices for publication of OGD can therefore be defined as the recommended ways of performing the tasks related to publication of Open Government Data.

Making data available online seems like a natural choice nowadays (Open Knowledge Foundation, © 2010-2012) and thus the OGD initiatives usually utilize the Web as a platform for publication and access to the open datasets. Therefore activities of the Word Wide Web Consortium4 are relevant to the OGD publication. Besides many technical specifications (e.g. XML5, RDF6 or DCAT7 to name a few) that help to build the interoperability, there are also relevant best practices issued or being developed by the W3C. Best Practices for Publishing Linked Data (Hyland et al., 2014) were published in January 2014. They specifically deal with publication of data that follows the Linked Data principles (see (Berners-Lee, 2006)). A more generic set of best practices for publication of data on the Web are currently being developed by the Data on the Web Best Practices Working Group (Łościo et al., 2015).

LAPSI project was aimed at dealing with the legal issues related to PSI and Open Data (KU Leuven, © 2010). PSI licensing is addressed by the Licensing Guidelines (De Rosnay et al., 2014) that builds upon the identified good practices for the license interoperability (Tsiavos, 2013).

Alongside the best practices targeting the technical and the legal aspect of OGD, various other issues are addressed by the best practices that are being developed in the Share-PSI project (Share-PSI 2.0,
Currently proposed best practices cover topics such as the cross-agency strategy for OGD, top management support, user engagement or the need for holistic metrics that would allow assessment of the value and impact of OGD or re-use of the PSI.

5. Open Government Data Publication Methodology

In this section MePOD-VS OGD publication methodology is described. The original version of this methodology is described in author’s thesis (Kučera, 2015). MePOD-VS methodology consists of a set of defined principles, roles and processes that are organized into groups/domains. For each of the proposed processes inputs and outputs are specified and for some of the processes the relevant standards were also identified.

Public administration in the Czech Republic is performed by various kinds of PSBs with different scope of authority and responsibility (Horzinková & Novotný, 2010). Besides this availability of the IT experts significantly differs between various kinds of the Czech PSBs (Czech Statistical Office, 2012). Lack of resources for OGD publication (especially in case of small agencies) is also one of the possible barriers identified by Janssen et al. (2012). Therefore the processes of the MePOD-VS methodology are classified according to the expected applicability of the processes in the context of different kinds of the Czech PSBs.

Berners-Lee (2006) suggested the 5-star deployment scheme for Open Data. In the MePOD-VS methodology this scheme is used to mark the relevant processes that are directly involved in publication of datasets on the particular level of the 5-star deployment scheme. This should ease identification of the minimum subset of processes that need to be implemented when for example Linked Open Data (5-star data) are published.

Discussion of the MePOD-VS methodology in its fullest extent is beyond the scope of this paper. Therefore the rest of the paper is focused on the process domains of the MePOD-VS methodology because these domains represent the key tasks of an OGD initiative. Model of the MePOD-VS process domains is depicted in figure 1.

![Fig. 1: Process domains of the MePOD-VS methodology, source: author](image)

In order to reflect some of the findings that resulted from the evaluation of the original version of the MePOD-VS methodology, model presented in figure 1 has been updated from the version presented in (Kučera, 2015). More specifically the scope of the methodology was extended to cover initiation and evaluation of an Open Data initiative, requirements analysis and goal setting was added as a separate domain and the scope of the user engagement process domain was extended (see below).
In terms of the MePOD-VS methodology an Open Data initiative is an initiative executed by some PSB. National or cross-agency OGD initiatives are not in the scope of the methodology.

Open Data initiative initiation might involve obtaining support of the top management of an organization, internal evangelization of the Open Data concept and securing participation of departments and other stakeholders. According to (Socrata, © 2015) obtaining support of the senior elected officials might help to achieve more results and to progress the initiative more quickly.

Requirements analysis and goal setting domain is aimed at understating the requirements of both the internal and external stakeholders and at setting the goals of the OGD initiative. According to (Socrata, © 2015) successful OGD initiatives are driven by clear and measurable goals that are aligned with the defined mission.

Open Data publication planning, Preparation of datasets and infrastructure, Open Data publication, cataloguing and maintenance and the Open Data archiving and retirement domains provide the necessary processes involved in the stages of the datasets lifecycle (for the lifecycle model see (Möller, 2013)). The main objective of the Open Data publication planning is to select a set of datasets for publication that is in line with the defined goals and to develop an Open Data publication plan that will be used to steer the OGD initiative. Datasets planned to be released need to be prepared, e.g. they might need to be transformed into a suitable machine-readable format, enriched with metadata and properly licensed. Once the datasets are prepared they need to be made accessible and discoverable (see for example (Lee et al., 2014)). Datasets and the respective metadata also need to be regularly updated (Dodds & Newman, 2015; Lee et al., 2014).

In order to be able to properly manage the end of the dataset lifecycle the Open Data archiving and retirement domain is a part of the MePOD-VS methodology. Relationship between the Open Data publication planning and the Open Data archiving and retirement indicates that in certain situations it might be necessary to add retirement of datasets as the planned activities. In the Czech Republic a new registry was established pursuant to the Act No. 304/2013 Coll. which also resulted in shift in responsibilities for collection of the relevant data from a ministry to the respective courts (Kubová, 2013). This example demonstrates that such changes in legislation might affect what datasets particular PSBs are able to publish as OGD.

Lee et al. (2014) view evaluating the progress of an OGD initiative and its impact as a critical element of its development and implementation. Mature organizations should also assess the progress against the defined Open Data strategy (Dodds & Newman, 2015). Therefore the Open Data initiative evaluation was added as a separate process domain into the MePOD-VS methodology that deals with the evaluation of progress against the Open Data publication plan and the defined goals.

User engagement and relationship management process domain is aimed at identification of both actual and potential users of the published data, assessment of the user demand and requirements and setting up and execution of the communication strategy. Its aim is also to ensure that it is possible to provide feedback on the published data. Lee et al. (2014) argue that Open Data need to be user-centric in order to bring social and economic benefits. User engagement had been part of the original version of the model in (Kucera, 2015), however in this updated model it is viewed as a crosscutting task that is performed throughout the whole OGD initiative rather than a step that follows the publication of datasets. While facilitation of the user feedback and re-use remains an important part of the OGD initiative8 this shift allows engaging users in the early stages of the OGD initiative which should help to establish a demand-driven release of data as suggested by Davies (2012). This in turn should lead to a better alignment of data demand and supply.

Besides the tasks of the domains depicted in figure 1 there are other activities that need to be performed during the OGD publication such as the data quality management, benefits management or risk management (Nečaský et al., 2014). These topics are covered by individual processes and not separate process domains in (Kucera, 2015). However according to the IT Governance Institute (2009) risk management should represent a continuous process. Therefore disciplines such as the risk management, benefits management or the data quality management might be related to more (if not all) process domains proposed in figure 1 in a way similar to the User engagement and relationship

8 Lack of feedback is viewed as barrier to the OGD re-use in (Janssen et al., 2012). Facilitating the feedback and reuse is seen as a recommended practice for example in (Lee et al., 2014; Ubaldi, 2013).
management process domain. Therefore the future research will focus on alignment of these disciplines with the newly redesigned process domains model.

Risk IT framework proposed in (IT Governance Institute, 2009) represents a generic IT risk management methodology. It is obvious that for some of the disciplines applied during the OGD publication, specific methodologies might be available. There are specific risks related to OGD publication (Kučera & Chlapek, 2014) but the future research should focus on alignment of the OGD publication methodologies with the relevant methodologies in other domains in order to avoid reinventing the proven practices.

6. Conclusions

Public sector bodies hold a significant amount of data that might be of potential interest to citizens and businesses. Data is a resource with a huge potential to re-use but this potential is yet largely untapped (G8, 2013). Machine-readability and legal openness are key attributes of Open Data that should make it easier to re-use. Therefore making the government data available as Open Government Data (OGD) should help to exploit its full re-use potential.

Public sector bodies (PSBs) face a number of challenges when publishing OGD (Janssen et al., 2012; Kučera et al., 2015; Ubaldi, 2013). As a response to the need to tackle these problems and challenges, OGD publication methodologies and best practices are being developed. In this paper emerging best practices for publication of OGD were introduced as well as the MePOD-VS OGD publication methodology.

The MePOD-VS methodology was originally proposed in (Kučera, 2015). In this paper updated model of the OGD process domains was presented that reflects some of the findings resulting from the evaluation of the original version of the MePOD-VS methodology. This model consist of the following process domains: Open Data initiative initiation, Requirements analysis and goal setting, Open Data publication planning, Preparation of datasets and infrastructure, Open Data publication, cataloguing and maintenance, Open Data archiving and retirement, Open Data initiative evaluation and the User engagement and relationship management. The proposed process domains are aimed at obtaining the top management support, setting of the goals of the OGD initiative as well as the evaluation of the OGD initiative. These process domains also contain processes involved in the dataset lifecycle from its beginning to its end. It is proposed to perform the user engagement throughout the OGD initiative which should ensure that the user-centric approach is applied and that the data release is demand-driven (see for example (Davies, 2012; Lee, et al., 2014)).

Future research should focus on alignment of the risk management, benefits management or the data quality management disciplines with the newly redesigned process domains model and on alignment of the OGD publication methodologies with relevant methodologies in other domains which might allow to build upon the proven practices during the OGD publication.

References


Cockburn, A., 2000: Selecting a Project’s Methodology, IEEE Software 17 (4), pp. 64-71

---

9 These disciplines are represented as cross-cutting activities in (Nečaský et al., 2014).

Davies, T., 2012: Supporting open data use through active engagement [Online]. Available at: http://www.w3.org/2012/06/pmod/pmod2012_submission_5.pdf [Accessed 2014-12-05]


IT Governance Institute, 2009: The Risk IT Framework. Rolling Meadows: IT Governance Institute


JEL Classification: O33, O38, O39, Z18