University of Economics, Prague
Faculty of informatics and statistics
Department of information technologies

Open Government Data

SUMMARY OF THE DOCTORAL DISSERTATION THESIS

Author : Ing. Jan Kučera

Thesis supervisor : prof. Ing. Jaroslav Jandoš, CSc.

Field of study : Informatics

jan.kucera [at] vse.cz
When citing please use the following reference:
Kučera, J.: Open Government Data: Summary of the doctoral dissertation thesis, VŠE-FIS, Prague, 2014

Prague, November, 2014

About this document

This document represents a summary of the research conducted as a part of the unfinished doctoral dissertation thesis *KUČERA, J.: Open Government Data, dissertation thesis, University of Economics, Prague*. The summary reflects the state of research as of November 2014. Therefore any research results presented in this summary are preliminary and it may change in the future.

Abstract

This Ph.D. thesis deals with Open Government Data and the methodology for publication of this kind of data. Public sector bodies hold a significant amount of data that can be reused in innovative way leading to development of new products and services. According to the Open Knowledge Foundation (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.*" Publication and reuse of Open Government Data can lead to benefits such as increased economic growth. State, society as well as the public sector bodies themselves can benefit from Open Government Data. However the public sector bodies currently face a number of problems and issues when publishing Open Government Data, e.g. regular updates of the published datasets is not always ensured. Different public sector bodies apply different approaches to publication of Open Government Data. The main goal of this thesis is to design the Open Government Data Publication Methodology which should address current problems related to the publication of Open Government Data.

Key words

Open Data, Open Government Data, Linked Open Data, LOD, OGD, levels of openness, public sector, data publication, data management.

Revision History

Revision	Date	Description	Author
1.0	2014-07-31	First version of the summary	Jan Kučera
1.1	2014-08-25	Correction of the text and figures	Jan Kučera
1.2	2014-10-04	Evaluation section added	Jan Kučera
1.3	2014-11-23	Changes in the thesis reflected	Jan Kučera
1.4	2014-12-08	Correction of the text and figures	Jan Kučera

Table of Contents

1 Introduction

This dissertation thesis deals with Open Government Data (OGD) and a methodology for its publishing. According to the Open Knowledge Foundation (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 Government Data has gained a lot of attention because it is viewed as one of the key aspects of the Open Government initiatives (Bauer & Kaltenböck, 2011). Although the idea of providing government data for re-use is older that the term Open (Government) Data, Open Government initiatives bring new emphasis on transparency an collaboration between government, citizens and business and on the free access to information. However OGD does not only support the Open Government agenda but OGD might be re-used to develop new products and services which might subsequently lead to the economic benefits as well (see for example (Buchholtz, et al., 2014) or (Vickery, 2011)).

OGD promises significant benefits to citizens, business and the public sector bodies as well. However the public sector bodies and users of OGD are currently facing a number of problems and issues when publishing and consuming OGD. Lack of standard process for OGD publication (Janssen & Zuiderwijk, 2012), unclear, missing or restrictive terms of use (Janssen, et al., 2012), lack of systematic measurement of the costs associated with the OGD publication (Ubaldi, 2013) or fact that the published open datasets are not always regularly updated (Tinholt, 2013) can be names as examples of the current problems and issues related to the OGD publication and consumption.

1.1 Research questions

Although the opening up data is sometimes presented as something that is easily achieved, according to (Janssen & Zuiderwijk, 2012) public servants who are in charge of the process of opening up data *„experience that opening might be more difficult than initially advocated."* This as well as the problems and issues related to OGD publication and consumtion might indicate that the process of OGD pubalication needs to be better understood in order to be able to effectivelly publish OGD that is easy to re-use by its potential users. Therefore the following research questions were formulated:

1. What tasks or processes needs to be performed when publishing OGD?
2. What roles and responsibilities should be defined in order to establish an organizational structure supporting OGD publication?
3. What are the differences between the various types of public sector bodies and how these differences might affect implementation of the OGD publication methodology?
4. How can the target level of data openness affect the OGD publication[1]?

1.2 Goals of the thesis

The main goal of this dissertation thesis is to design a methodology for publication of Open Government Data. Such methodology should describe recommended OGD publication processes as well as a set of roles and responsibilities that will support the OGD publication. It can also provide guidelines and recommendations that might help the OGD publisher to deal with commonly faced problems and issues.

[1] 5 star schema proposed by Sir Tim Berners-Lee (2006) represent a model of levels of data openness. In general openness level represents a set of requirements that an open dataset should conform to.

The main goal and the supplementary goals of this dissertation thesis are summarised in the table 1.

Table 1: Goals of the dissertation thesis, source: author

ID	Goal	Type
MG1	Design a methodology for publication of Open Government Data	Main
SG1	Design the OGD publication methodology so that different levels of data openness are reflected in the recommendations.	Suppl.
SG2	Classify the Czech public sector bodies into categories relevant to the OGD publication.	Suppl.
SG3	Develop a model of roles participating on the OGD publication.	Suppl.
SG4	Propose the implementation process of the methodology suitable for a selected subset of Czech public sector bodies.	Suppl.
SG5	Evaluate the designed methodology.	Suppl.

2 Research approach

The main goal of this thesis is to design a methodology supporting the OGD publication that would provide recommended OGD publication process, organizational structure as well as it should address the problems and issues faced by the OGD publishers. In order to achieve this goal the Design Science approach is applied in this thesis, especially the guidelines for Design Science in information systems research proposed by Hevner et al. (2004). Research methodology of this thesis is based on the *Design Science Research Methodology for Information Systems Research (DSRM)* proposed by Peffers et al. (2007). Figure 1 depicts the research methodology applied in this thesis.

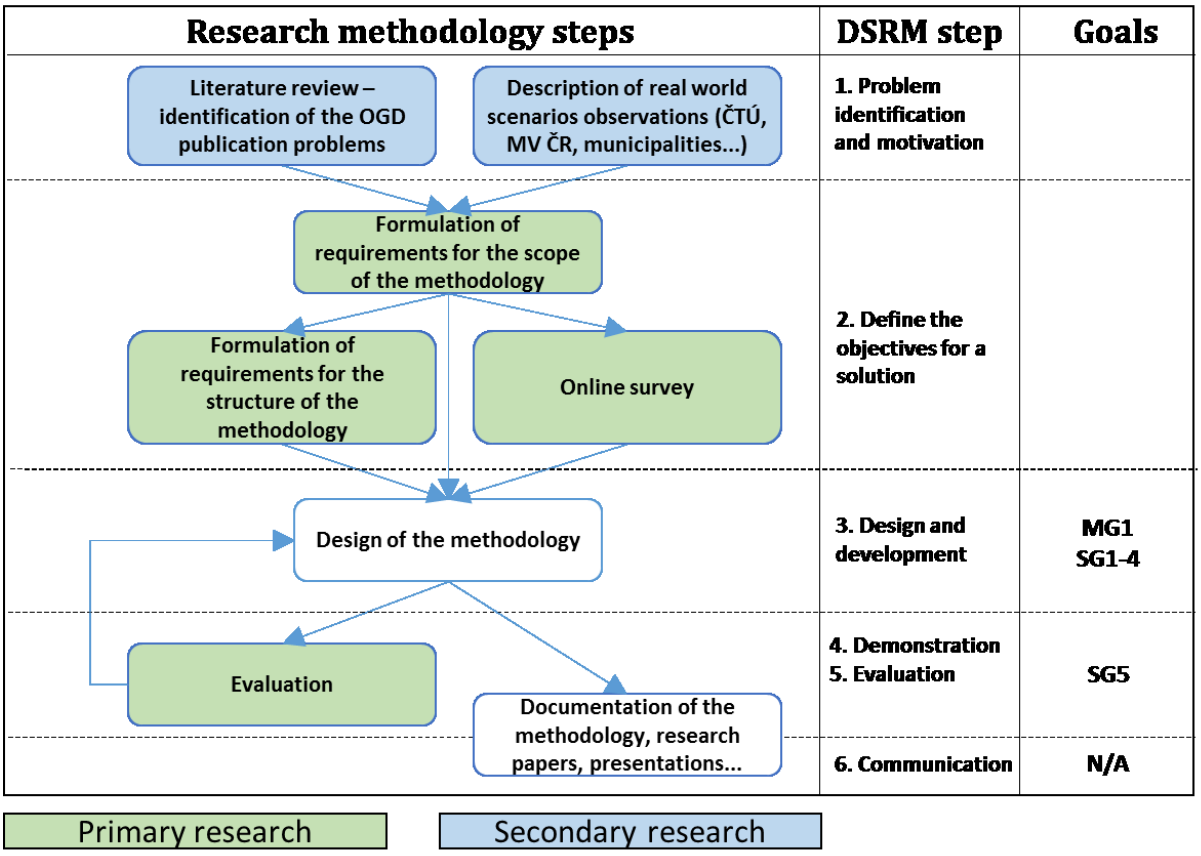

Figure 1: Research approach applied in the thesis, source: author

3 Open Government Data Publication Methodology

Artefact that is being designed in this thesis is the OGD publication methodology called MePOD-VS[2]. According to Buchalcevová (2009) a methodology is "*a set of methods and procedures for performing a certain task.*" In the thesis the OGD publication methodology is defined as a set of methods, procedures 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 and how to perform the steps of the OGD publication process.

Based on the analysis of problems and issues related to the OGD publication and consumption and analysis of the real world scenario observations a set of requirements for the scope of the methodology were formulated, i.e. topics that should be addressed in the methodology. These requirements are described for example in a research report (Kučera, 2014).

Methodological elements or building blocks that currently make up the MePOD-VS methodology are:

- domain,
- process,
- principle,
- role and its competencies,
- type of public sector body[3] and assessment of suitability of a particular process for that type of a public sector body,
- standard and recommendation for its application,
- artefact (input or output of a process),
- template of an artefact,
- openness level model, level of openness that is a part of some openness level model and the requirement associated with a certain level of openness and
- software tool.

Based on the identified requirements a set processes was designed through which the OGD publication is realized. These processes form six process domains as depicted on figure 2.

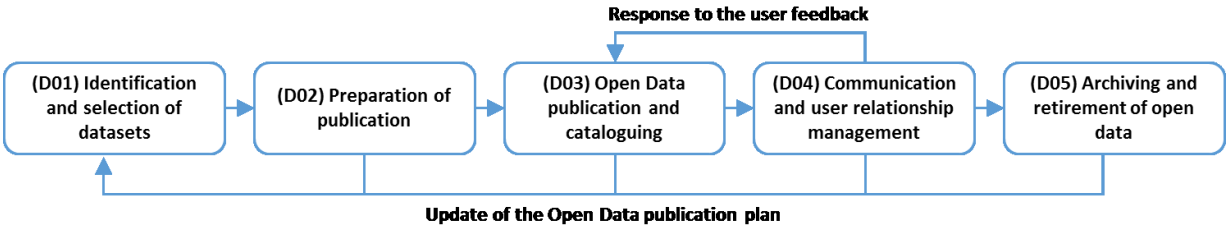

Figure 2: OGD publication process domains, source: author

In the MePOD-VS methodology the OGD publication starts with the identification and selection of datasets for opening up (*D01*). The goal of this process domain is to analyse the datasets held by the public sector body and identify suitable datasets that can be published as Open Data according to the results of the analysis. Basic attributes of the datasets like the format the data should be published in or the update frequency should be determined at this stage as well. Analysis performed during the identification of datasets should involve risks and benefits analysis and analysis of impact on the

[2] This is an abbreviation of the Czech title "**ME**todika **P**ublikace **O**tevřených **D**at **V**eřejné **S**právy."
[3] Due to the differences in organization of the public sector in different countries, the classification reflects the public sector in the Czech Republic only.

information systems. Estimation of effort needed to publish the datasets should be performed as well. Based on these information datasets for opening up should be prioritized and the Open Data publication plan should be developed.

The goal of the process domain (*D02*) *Preparation of publication* is to prepare the datasets so that they conform to the defined technical and legal openness requirements. Processes of this domain aim mainly at the design of the target schema of the data and the related design of the automated transformation procedures (ETL), preparation of metadata, implementation of the infrastructure supporting the OGD publication including the data catalogue, selection of the appropriate open licence, development of the communication strategy and the data quality assurance strategy.

The goal of the process domain (*D03*) *Open Data publication and cataloguing* is to make the selected open datasets as well as the related metadata available for reuse and to ensure their regular maintenance and updates. This domain is also aimed at ensuring the required level of data quality of the published data and metadata.

Processes of the domain (*D04*) *Communication and user relationship management* ensure that the users are informed about the available open datasets by its publisher and the public sector body is able to receive and respond to the feedback provided by the users. Regular assessment of the demand for data is a part of this domain. Results of this assessment can initiate an update to the Open Data publication plan and extension of the list of datasets that are planned to be opened up.

Feedback provided by the users might also help to identify issues in the data and the way it is being published and thus it can help to improve quality of the data and the OGD publication practices. Processes of the domains *D03* and *D04* should not be isolated but the feedback should be used as an input to the maintenance of the data. Updates of the data and metadata can in turn lead to the new user feedback. Processes of these domains can be executed more than once during the implementation of the Open Data publication plan.

Long term availability of the published data should be ensured in order to make the datasets reliable data sources for users and applications. However in certain situations maintenance or even the publication of some datasets might be terminated. For example if an open datasets is created by transformation of some primary data that is being collected pursuant to some legal act, amendments of this act or its repeal might lead to termination of collection of the primary data and as a result the related open dataset will not be updated anymore. Datasets that have been already published might be kept available to the users unless it violates the newly effective legislation. In situations when the maintenance or publication of some datasets is terminated users should be informed about the changes in status of the datasets and the metadata should be updated accordingly. Therefore the goal of the processes in the (*D05*) *Archiving and retirement of open data* domain is to manage the end of the open datasets life-cycle.

Publication of Open Data is not seen as a one-time activity of the publisher but as a stable function of an organization. Therefore the processes in the process domains described above should be continuously repeated. New iteration over the processes might be triggered for example by the periodic assessment of the Open Data publication plan or by its ad hoc update as a response to the immediate needs or events.

To support the OGD publication a set of roles was defined and responsibilities of were set. Prior research on this topic (Kučera, et al., 2013) as well as analysis of roles described in other

methodologies ((Chlapek, et al., 2012), (Socrata, 2013)) served as a basis for designing the set of roles used in the MePOD-VS methodology. The following roles are involved:

- (R01) Data publisher
- (R02) OGD coordinator
- (R03) Data curator
- (R04) IT professional
- (R05) Legal expert
- (R06) Communication expert
- (R07) Data catalogue owner
- (R08) Data catalogue provider
- (R09) Website owner
- (R10) Website provider
- (R11) Catalogue editor
- (R12) Ontology designer
- (R13) Linked Data expert
- (R14) End user

4 Expected benefits

Answering the research questions and designing the MePOD-VS methodology should provide the following benefits:

- Identification and description of the process domains supporting the OGD publication – description of the OGD publication process domains that also reflects the end-of-life phase of the dataset lifecycle should help to guide the activities related to the publication of OGD.
- Formulation of OGD publication methodology requirements – current problems and issues related to OGD publication were identified and based on analysis of these problems a set of OGD publication methodology requirements were formulated. These requirements might be used for comparison of existing methodologies as well as they might be used in development of other OGD publication methodologies. The MePOD-VS methodology should be designed so that it helps the OGD publishers to deal with the identified problems. However other approaches and solutions to the identified problems might be proposed by other researchers and practitioner. Therefore discussion of the current problems related to the OGD publication might also stimulate new research.
- MePOD-VS should focus on some aspects that are only partially addressed in the available OGD publication methodologies like risk analysis or long term maintenance of the open datasets. Therefore it should help the OGD publishers to better manage the OGD publication.

5 Evaluation

Results of the analysis of the current problems and issues related to the OGD publication and consumption, the set of requirements for the scope of the OGD publication methodology formulated in the thesis as well as some attributes of the MePOD-VS methodology were evaluated in the

COMSODE[4] project during the development of the deliverable D5.1 "*Methodology for publishing datasets as open data*" (Nečaský, et al., 2014). In July 2014 the MePOD-VS methodology as well as the underlying analysis of the OGD publication problems were taken as an input into the development of the OGD publication methodology developed in the COMSODE project. Requirements for the OGD publication methodology formulated in this thesis were added to the specific requirements described in the Description of Work of the COMSODE project and based on this set of requirements the COMSODE OGD publication methodology was developed. Although the methodology developed in the COMSODE project differs from the MePOD-VS methodology, these methodologies share some common aspects like focus on tasks performed in the phases of the OGD life-cycle, description of responsibilities of the defined set of roles or identification and description of artefacts that serve as inputs and outputs of the tasks through which the OGD publication is performed. The COMSODE OGD publication methodology itself was reviewed by the members of the COMSODE project User Board. Based on the recommendations of the reviewers the COMSODE OGD publication methodology was refined, however some of the recommendations are applicable to the MePOD-VS methodology as well and thus they will be used to improve the quality of the MePOD-VS methodology.

The MePOD-VS methodology will be further evaluated by interviews with the OGD experts. The objective is to evaluate that the methodology focuses on appropriate domains and it should be also discussed what the preferred way of handling the user feedback is. Another objective of the interviews is to identify what outputs of the processes should be described in the methodology, to review and evaluate the defined set of roles and their responsibilities and to discuss the differences between the various types of the public sector bodies and how these differences could affect the OGD publication.

The target group of OGD experts should consist of representatives of different types of public sector bodies as well as researchers, consultants or relevant representatives of NGOs.

The MePOD-VS methodology should be also evaluated in a study focused on the implementation of the MePOD-VS methodology in at least one Czech public sector bodies.

6 Conclusions

Public sector bodies hold significant amount of valuable data than can be reused in new and innovative applications and services. OGD promises significant benefits to citizens, business and the public sector bodies as well. However the public sector bodies are currently facing a number of problems and issues when publishing OGD (see (Kučera, 2014)). Research outlined in this paper aims at identification of the current problems and issues related to OGD publication and consumption and at designing the OGD publication methodology that would help the OGD providers to better manage the OGD publication and to deal with some of the current problems.

Methodology MePOD-VS introduced in this paper proposes five process domains through which the OGD publication is realized: (*D01*) *Identification and selection of datasets*, (*D02*) *Preparation of publication*, (*D03*) *Open Data publication and cataloguing*, (*D04*) *Communication and user relationship management* and (*D05*) *Archiving and retirement of open data*. In order to provide organizational structure supporting the OGD publication, MePOD-VS methodology proposes a set of

[4] http://www.comsode.eu/

roles and sets their responsibilities for the defined processes. Alongside the phases, processes, roles and responsibilities, other methodological elements are used to provide additional guidelines and recommendations.

7 References

Bauer, F. & Kaltenböck, M., 2011. *Linked Open Data: The Essentials.* Vienna, Austria: edition mono/monochrom.

Berners-Lee, T., 2006. *Linked Data - Design Issues.* [Online] Available at: http://www.w3.org/DesignIssues/LinkedData.html
[Accessed 2012-02-20].

Buchalcevová, A., 2009. *Metodiky budování informačních systémů.* Praha: Nakladatelství Oeconomica.

Buchholtz, S., Bukowski, M. & Śniegocki, A., 2014. *Big and open data in Europe. A growth engine or a missed opportunity?.* [Online] Available at: http://www.bigopendata.eu/wp-content/uploads/2014/01/bod_europe_2020_full_report_singlepage.pdf
[Accessed 2014-07-10].

Hevner, A. R., March, S. T., Park, J. & Ram, S., 2004. Design Science in Information Systems Research. *MIS Quarterly*, 28(1), pp. 75-105.

Chlapek, D., Kučera, J. & Nečaský, M., 2012. *Metodika publikace otevřených dat veřejné správy ČR.* [Online] Available at: http://www.mvcr.cz/soubor/metodika-publ-opendata-verze-1-0-pdf.aspx
[Accessed 2014-07-29].

Janssen, M., Charalabidis, Y. & Zuiderwijk, A., 2012. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4), pp. 258-268.

Janssen, M. & Zuiderwijk, A., 2012. Open data and transformational government. *Proceedings of the Transforming Government Workshop 2012 (tGov2012)*.

Kučera, J., 2014. *Methodologies for publication of Open Government Data.* [Online] Available at: http://nb.vse.cz/~xkucj30/dissertation/Kucera_OGD_methodologies_EN_v1.pdf
[Accessed 2014-07-30].

Kučera, J., Chlapek, D. & Nečaský, M., 2013. Linked Open Data Stakeholder Roles. *CONFENIS 2013*, p. 11–28.

Nečaský, M. a další, 2014. *Deliverable D5.1: Methodology for publishing datasets as open data.* [Online] Available at: http://www.comsode.eu/wp-content/uploads/D5.1-Methodology_for_publishing_datasets_as_open_data.pdf
[Accessed 2014-09-10].

Open Knowledge Foundation, 2012. *The Open Data Handbook.* [Online]
Available at: http://opendatahandbook.org/
[Accessed 2012-08-30].

Peffers, K., Tuunanen, T., Rothenberger, M. & Chatterjee, S., 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), pp. 45-77.

Socrata, 2013. *The Open Data "A Team" Workbook.* [Online] Available at: http://www.socrata.com/wp-content/uploads/2013/02/Chapter-3-The-Open-Data-A-Team-Workbook.xlsx
[Accessed 2014-02-15].

Tinholt, D., 2013. *The Open Data Economy. Unlocking Economic Value by Opening Government and Public Data.* [Online] Available at: https://www.capgemini-consulting.com/ebook/The-Open-Data-Economy/files/assets/downloads/publication.pdf
[Accessed 2013-02-25].

Ubaldi, B., 2013. *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives.* OECD Publishing.

Vickery, G., 2011. *Review of recent studies on PSI re-use and related market developments.* [Online] Available at: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/psi_final_version_formatted.docx
[Accessed 2012-12-29].