Towards a Methodology of Multi-criteria Prioritization of Open Data for Public Release

Working Paper

Alexei Botchkarev

Abstract

Open Data movement is gaining momentum around the world as governments are striving to improve transparency, accountability and public engagement. Huge contents of data repositories, accumulated by public services and previously restricted from general public access, need to be reviewed, prepared and openly published to meet expectations of the citizens. Governments are facing challenges in determining how datasets should be prioritized for public consumption.

The results of this study are intended to alleviate existing problem of dataset prioritization. An Open Data prioritization methodology has been developed including criteria, model and analytics. The proposed multi-criteria prioritization methodology is value-focused and is based on the calculation of the net value (i.e. value minus cost) of the dataset adjusted for risk and urgency. The methodology has been operationalized through the quantification of the model allowing transforming mostly qualitative evaluators' assessments into numbers and offering analytics to come up with a priority score of each dataset. Practical contribution of the study is in offering simple, transparent and reliable analytics to facilitate decision-making process of prioritizing open datasets for public release.

Keywords: open data, multi-criteria, prioritization, public release, methodology

Introduction

Open Data movement is gaining momentum striving to improve governments' transparency, accountability and public engagement. An ever increasing number of governments are cooperating within such communities as International Open Data Charter (www.opendatacharter.net), Open Government Partnership (www.opengovpartnership.org), Open Definition (www.opendefinition.org), Open Knowledge International (www.okfn.org), etc.

As a tangible outcome of the Open Data movement, various governments published thousands of datasets which were previously restricted from general public access. For example, large open data repositories were created by the governments of Canada (http://open.canada.ca/data/en/dataset), US (www.data.gov), UK (www.data.gov.uk) facilitating faster and easier access to government data.

Theoretical issues of open data, such as their benefits and challenges, potential impacts on the ability of making evidence-based decisions, etc., have been tackled in a number of academic articles, e.g. (Davies & Bawa, 2012; Davies & Perini, 2016; Larquemin, Buteau & Mukhopadhyay, 2016).

Issues of practical implementation of the open data strategies received less attention in academic literature. This article aims to add to this body of knowledge and examines one of important implementation issues facing public service practitioners: how to prioritize open datasets for public release? The importance and timeliness of the issue is determined by several factors. First, any government entity collects and maintains large number of datasets (normally, hundreds or thousands, as it is evidenced by the websites cited above). Preparing them for public release can be an overwhelming task, especially at the initial stages, as more and more governments are embarking on this journey. Second, limited human and other types of government resources are not sufficient to evaluate, review, prepare and publish all datasets in a reasonably short period of time. Certain queue of datasets in the publication "pipeline" is almost inevitable. Finally, there are optimistic expectations and impatience on behalf of the different groups of potential users to get access to the datasets they need. Making suboptimal decisions on the sequence of opening datasets for public consumption may result in perceptions of inefficient government service. Existing relevant literature on the topic is scarce, e.g. (Data.gov Interim, 2010; Federal CIO, 2015).

The purpose of the study is to develop an open data prioritization methodology including objectives of selecting a set of prioritization criteria, and building mathematical model.

Several methodologies were used to achieve the research objective: identification of related peer-reviewed papers, critical literature review, critical thinking, inductive reasoning. The considerations of prioritization are generic and applicable to any field (i.e. government ministry or agency).

The results of this study are intended for public service practitioners, business, data and information analysts, information systems, program and policy evaluators, project and program managers, and researchers interested in open data, technology solutions and business management.

The paper is structured as follows. Next section examines prioritization criteria and identifies subcriteria that can be used to evaluate datasets. It is followed by a section in which the author develops prioritization model and derives qualitative relations to incorporate criteria assessments. Then, the model is quantified through the introduction of mathematical formulas allowing calculating priority scores to the evaluated datasets. Finally, some key points of the methodology implementation are highlighted and concluding remarks are presented.

Prioritization purpose and criteria

The prime purpose of the prioritization is in ensuring an expedited release of data with high social, health and economic value. Prioritization methodology will enable dataset owners to evaluate which datasets are most critical to share with public audiences and provide a structured process of ranking the datasets based on clear criteria. Results of prioritization will inform dataset owners in assigning resources (human and other) and scheduling publication of datasets.

It should be noted that prioritization is only one of several open data management processes which also include identification of datasets, building inventories, publishing datasets, refreshing datasets, etc. Prioritization is not overlapping or substituting any of these processes, e.g. prioritization is not intended to answer the question whether a dataset can be publicly released or not (e.g. for security or privacy reasons) – this question must be answered within the context of other open data management processes.

Effective prioritization requires comprehensive evaluation of the datasets from several perspectives making the task a multi-criteria decision making exercise. A literature review has been conducted to identify and select a set of criteria used for prioritization. The search was performed in a broader area which included criteria selection and prioritization efforts in project management, e.g. project prioritization, portfolio prioritization, program evaluation (Data.gov Interim, 2010; Federal CIO, 2015; Padovani, Muscat, Camanho & Carvalho, 2008; Benedetto & van der Linden, 2015; Posavac, 2015). Critical analysis of the literature review findings allowed selecting the following set of criteria for the proposed methodology:

- Value – a measure of expected positive impact of releasing data;
- Urgency – a measure of necessity of expeditious data release;
- Risk – a measure of expected exposure to threats and unintended consequences associated with releasing data;
- Cost – a measure of expected expenditure and efforts to prepare data for release and maintain the data once made public.

Single-word title of each criterion must be understood in the broadest meaning. The meaning and context of the criteria are identified through the introduction of the sub-criteria in the form of questions. Later, questions are combined into a questionnaire.

The value criterion may have at least two aspects: the impact on government efficiencies and impact on public consumers or society at large. The following types of questions (sub-criteria) can be considered when determining the value of releasing a dataset: Has the data collection or production been mandated by a legal act or statute? Does the dataset span interests of multiple sectors of economy?

The urgency criterion may be considered in relation with certain timelines or deadlines. For example, a hackathon has been scheduled and participants would need input data. These data should be released to public prior to the hackathon date. The following types of questions can be considered when determining the urgency of releasing a dataset: Are there imminent deadlines for activities that will utilize the dataset? Will the data release bring immediate value to the public?

The risk criterion is intended to evaluate potential negative consequences associated with sharing the dataset. The following types of questions can be considered when determining the risk of releasing a dataset: Are there any potential issues with data credibility or validity? Are there risks of data being misinterpreted? The assumption is that initial risks of breaching privacy or security

have been evaluated at previous steps of the open data management. Prioritization deals with residual risks.

Cost criterion is intended to determine all types of associated costs including direct and indirect costs, e.g. salaries and wages plus benefits for full time equivalent positions and consultants, hosting expenses, etc. The following types of questions can be considered when determining the cost of releasing a dataset: What is the estimated overall cost/effort for data preparation? What is the return on investment (ROI) of the dataset release?

Each entity responsible for publishing open data should develop an individual questionnaire to facilitate evaluation of the datasets. The questionnaire should contain sets of questions (subcriteria) and prompts that would guide evaluators through selected criteria to an objective ranking of the datasets. The number of questions included in the questionnaire is specific to the needs of each organization and is a matter of discretion. Obviously, fewer questions will save time and workload for evaluators. But, stretching it to the limit, it is clear that having one question per criterion is not enough, e.g. if a single question is "what is the value of this dataset?" – most likely each evaluator would understand the meaning of the question in his/her own way and evaluation will not be consistent. As we mentioned, value has several aspects, e.g. for value for the government efficiency, public, research organizations, etc. An important purpose of the questions and prompts is to provide guidance and context for evaluators. It appears that four – five for each criterion is a reasonable number of questions in a questionnaire. Such document can be completed within five - ten minutes by a subject matter expert (SME).

Prioritization model

All four criteria, i.e. value, cost, risk and urgency, must be used integratively in determining the priority score of each dataset. Analytics developed for the multi-criteria prioritization is based on the following contemplations and dependences. The main focus in determining the priority is on the value of the dataset for public use or government efficiencies. Value may be decreased, if the cost of publishing and maintaining the dataset for public is tangible. The value may be further lowered because of the risks associated with making dataset public. If publication of the dataset is related to some events that require urgent actions, the priority of this dataset increases. Based on the explained approach, the priority scoring is calculated according the following formula:

$$PriorityScore = (Value - Cost) \times (1 - Risk) + Urgency \quad (1)$$

Using this formula, each dataset on the prioritization list will be assigned a priority score. Quantification of the criteria and the process of prioritization are provided in the next section.
The proposed multi-criteria prioritization framework is value-focused and is based on the calculation of the net value (i.e. value minus cost) of the dataset adjusted for risk and urgency.

Model quantification

To operationalize proposed prioritization model we need to develop the process of model quantification (i.e. how to transform mostly qualitative assessments into numbers and how to use

these numbers to come up with a priority score of each dataset). Quantification of the model includes three steps:

- Setting weights to sub-criteria (i.e. each question of the questionnaire).
- Quantifying criteria assessments.
- Quantifying priority scores.

Setting weights to the sub-criteria

In an earlier section, we identified four criteria that are used to evaluate priority of the datasets. For each criterion, we defined sub-criteria (presented as questions) that will help making evaluation process more structured. Analysis of the sub-criteria has shown that each of them has different importance (weight). For example, when we determine value of the dataset, the answer to the question "Has the data collection or production been mandated by an act or statute?" is understandably more important than the answer to the question if "free tools are available to the public to manage a dataset". This should be taken into account during evaluation. Several SMEs may be called upon to assess relative weights of the sub-criteria. Each sub-criterion can be assigned a quantitative weight: a number in the range from 10 (very important question) to 1 (least important question). SMEs can be asked to evaluate sub-criteria independently, and then the overall weight for each question will be calculated as an average. It should be noted that the weights (as well as the questions) are organization-specific and reflect the objectives of the organization in the current envinronment.

Quantifying criteria assessments

Evaluation of an individual dataset involves assessment of the dataset by each of the sub-criteria (i.e. answering all questions of the questionnaire). The easiest approach would be to give Yes or No answers and assign certain quantitative levels to each type of answer. However, some questions do not have clear Yes-No answers, or a SME who is making an assessment may not have complete information for a definitive response. To accommodate such situations, a numeric Likert-type rating scale can be used. To perform an assessment of a dataset, SME will be selecting one of five (5) response options to each of the questions (sub-criteria): 1. Strongly disagree; 2. Disagree; 3. Neutral; 4. Agree; 5. Strongly agree.

The number of the response option will be stored as a rating value for the question. Some of the response options can be given additional prompts to facilitate evaluations.

The score for each criterion is calculated as a sum of all scores of this criterion questions taking into account sub-criteria weights.

For example, Value score is calculated as:

$$\boldsymbol{Value_i = \sum_{n=1}^{N} VR_n \times VW_n} \qquad (2)$$

where VR_n is the rating of the n-th question (sub-criteria) assigned by SME;

VW_n is the weight of the n-th question;

N is the number of questions (sub-criteria) for the Value criterion in the questionnaire.

Quantifying priority scores

Initially calculated criteria scores (e.g. $Value_i$, etc.) do not reflect relative importance/weights of the criteria and must be adjusted. Criteria score ranges adjustments have been based on the following contemplations:

- Dataset Value is the key criterion and should be given significant weight.
- The amount of the Cost score should be lower than Value to avoid negative net value result. This point reflects the notion that publishing data has value for the public and government even if the cost of the process is high.
- Urgency criterion should be at the Cost score level.
- Risks are usually measured as a probability of occurrence of a negative event. In most situations risks are inevitable, so the situations with zero risk or probability of risk equals one are unlikely.

The following adjusted score ranges may be proposed (see Table 1).

Table 1: Adjusted criteria score ranges

	Value	**Cost**	**Risk**	**Urgency**
Maximum	120	20	0.75	15
Minimum	60	7	0.25	0

Mapping of the criteria score calculated within the initial range to the adjusted score range was performed using linear mapping formula (Heidari, Heidari & Homaei, 2014; Sun, Peng, Chen & Shukla, 2003):

$$\boldsymbol{Value} = (\boldsymbol{Value_i} - \boldsymbol{VImin}) \times (\boldsymbol{VAmax} - \boldsymbol{VAmin}) \,/\, (\boldsymbol{VImax} - \boldsymbol{VImin})) + \boldsymbol{VAmin} \quad (3)$$

where, $Value$ – adjusted value criteria score; $Value_i$ – initially calculated criteria score with formula (2);
VImin – initial minimum level;
VImax - initial maximum level;
VAmin - adjusted minimum level;
VAmax – adjusted maximum level.
Similar formulas should be used for the other criteria.

Formula (3) uses initial minimum and maximum values of the criteria, i.e. criteria score ranges. These amounts are organization-specific and are determined by the design of the evaluation questionnaire. For each criterion, these amounts depend on the number of the sub-criteria in the

questionnaire and their weights. They are calculated by using formula (2) substituting VR_n = 1 for initial minimum and VR_n= 5 for initial maximum.

Priority score for each dataset is calculated using formula (1) while substituting results of calculations for each criterion by formula (3). The values of the final priority score will be in a convenient range [10, 100]. The higher priority score signifies higher social, health and economic importance of the dataset and suggests the need for its expedited release to public. All organization's datasets are to be set in a queue for public release according to their priority scores.

Conclusion

An Open Data prioritization methodology has been developed including criteria, model and analytics. The proposed multi-criteria prioritization methodology is value-focused and is based on the calculation of the net value (i.e. value minus cost) of the dataset adjusted for risk and urgency. The methodology has been operationalized through the quantification of the model allowing transforming mostly qualitative evaluators' assessments into numbers and offering analytics to come up with a priority score of each dataset. Practical contribution of the study is in offering a simple, transparent and reliable model to facilitate decision-making process of prioritizing open datasets for public release.

References

Benedetto, H., & van der Linden, J. (2015). Criteria Definition for the Selection of Strategic Design Projects in Product Development Companies. *The Journal of Modern Project Management*, *3*(1).

Data.gov Interim Identification & Prioritization Process and Guidelines v1.0. (2010). US Department of Transportation. Available at: https://www.transportation.gov/sites/dot.dev/files/docs/identpriorguidelines1.0.pdf

Davies, T. G., & Bawa, Z. A. (2012). The promises and perils of open government data (OGD). *The Journal of Community Informatics*, *8*(2).

Davies, T., & Perini, F. (2016). Researching the emerging impacts of open data: revisiting the ODDC conceptual framework. *The Journal of Community Informatics*, *12*(2).

Federal CIO. (2015). Open Data Prioritization Toolkit. Federal CIO Council Innovation Committee. June 2015. Available at: https://cio.gov/wp-content/uploads/filebase/cio_document_library/Open%20Data%20Prioritization%20Toolkit_Summary.pdf

Heidari, M., Heidari, A., & Homaei, H. (2014). Analysis of pull-in instability of geometrically nonlinear microbeam using radial basis artificial neural network based on couple stress theory. *Computational intelligence and neuroscience*, 2014, 4.

Larquemin, A., Buteau, S., Mukhopadhyay, J.P. (2016). Open Government Data and Evidence-based socio-economic policy research in India: an overview. *The Journal of Community Informatics, 12*(2), (Special issue on Open Data for Social change and Sustainable Development), 120-147.

Padovani, M., Muscat, A. R. N., Camanho, R., & Carvalho, M. D. (2008). Looking for the right criteria to define projects portfolio: multiple case study analysis. *Product: Management & Development*, *6*(2), 127-134.

Posavac, E. (2015). *Program evaluation: Methods and case studies*. Routledge.

Sun, Y., Peng, Y., Chen, Y., & Shukla, A. J. (2003). Application of artificial neural networks in the design of controlled release drug delivery systems. *Advanced Drug Delivery Reviews*, *55*(9), 1201-1215.