Pathways for creating value with open data

Rafik Chelah - 4113136
Faculty of Technology, Policy & Management, Delft University of Technology
Jaffalaan 5, 2628 BX Delft, the Netherlands
r.chelah@student.tudelft.nl

ABSTRACT

Open government data (OGD) has much potential, however, its usage lags behind. It is often time consuming and difficult to determine the activities needed to use open data in general. The aim of this research is to develop pathways to bridge the gap between free available open data and its usage by end-users. Pathways should enable end-users to take advantage from open data by structuring the process of open data usage. Hereby, user goals are defined in terms of research questions and/or hypotheses. Before the develop pathways, a literature review is conducted in order to identify functions that are required for processing the effective use of open data. The functions are presumed to be available on an open data portal, including build-in techniques that enable online open data analyses. Each pathway represents a different approach and therefore strategy to make use of open data on a valuable manner. The pathways provide a sequence of steps as a guidance for end-users to process raw datasets into valuable insights. Within the pathway each step represent tasks that need to be fulfilled to complete the data analysis project.

Consequently, the pathways are, evaluated and improved by collecting feedback after interviewing five respondents with differing expertise. The advantage of pathways is that end-users can focus more on a specific domain problem through guided tasks as part of each step in a pathway to process the data accordingly. Pathways can save time by already providing standard steps that are needed to prepare and analyze open data. Pathways provide an overview of what end-users are capable to do with open data beforehand. The advantages of pathways include time efficiency, effectiveness of approach, motivation, inspiration, and recognition of sequential logic in the complex notion of open data usage. However, different end-users have different goals and therefore different user needs. In order to improve the pathways, more research is needed to decide what the user needs are.

Index terms→**Social and professional topics→Professional topics→ Computing and business→Socio-technical systems.**

Keywords

Open data; portal; open government data; portal functions; pathways

1. INTRODUCTION

Open data is valuable because of new developed technologies that enable the development of new services, which have attracted the attention of governments (Huijboom & Van den Broek, 2011). The issue is that open data only becomes valuable in used cases instead of considering open data as valuable itself (Janssen et al., 2012). Nowadays more and more data has become available (Chen et al., 2012). Despite its availability the use of the data lags behind and there are many barriers (Zuiderwijk et al., 2014a). The users of open government data (OGD) should be enabled to understand data to create value from them.

The definition of OGD is a combination of government data and open data, which can be defined as follows: Government data is ***"any data and information produced or commissioned by public bodies" (Ubaldi, 2013, p. 6)***, whereas Open Data can be defined as ***"data that can be freely used, re-used and distributed by anyone, only subject to (at the most) the requirement that users attribute the data and that they make their work available to be shared as well" (Ubaldi, 2013, p. 6)***.

There are high expectations for OGD, but meanwhile the user perspective on open data remains unclear for governments (Zuiderwijk et al., 2014a). The acknowledgement of OGD value is not enough, although the use of it, in combination with social intelligence that users have in terms of unique knowledge, can be valuable (Janssen et al., 2012). Possible barriers that conflict this, should be mitigated (Nugroho, 2013).

One of the important reasons that OGD still has unfulfilled potential is that its usage is difficult due to the restrictive user conditions (de Rosnay et al., 2014). Pathways that can mitigate the difficulties in using open data do not exist in research yet. Definition for a pathway is as follows:

"A way of achieving a specified result; a course of action" (Oxford, 2018: 1).

Because there are no pathways for the use of open data yet, in this article we introduce pathways. The development of pathways represents hereby the scientific contribution of this paper. Different types of user goals can be linked to different pathways to achieve those user goals. To support value creation with open data, the user

goals are structured by means of pathways. The scope of this article is therefore limited to end-users of open data who are interested in OGD, possibly in combination with other open data sources.
Four different pathways are defined to achieve user goals in a beneficial and effective manner from perspective of the end-user. The pathways are presented in form of a sequential range of steps to be followed in chronological order by each user. Because this is an initial exploration of pathway development, the first four pathways are developed as an impetus for more pathways. These pathways are consequently evaluated by demonstrating them to five interviewed experts with different expertise. The received feedback after the interviews is used to adjust and improve the pathways. In the future more different and possible adjusted versions of pathways can be developed, if there is demand for it. The latter can be part of future research to discover user needs in the field of open data analyses.
The outline for the rest of the article is represented in sections. Section 2 presents optional functions for an OGD portal, as well as suggestions for improving the user friendliness of OGD on European level. The functions for processing data analyses tasks, are taken into account for the design of pathways. Section 3 explains the research method by means of an information systems research framework as the overview of raw research material. The research method is in combination with design science steps that are used in this article as the basis for pathway development. Section 4 describes suggestions to mitigate the differences between open data offerings. Furthermore, the pathways are initiated. Consequently, section 5 presents an evaluated of the pathways by interviewing five expert on the open data field. Finally, section 6 presents the conclusion with suggestions for future research.

2. Functions for processing data

The goal in this section is to provide input information for the development of pathways. After a literature research several functions are found. A selection of these functions are mentioned to show what is possible in the field of open data processing via an open data portal. The functions that are mentioned in literature for an OGD portal are as follows (Zuiderwijk, 2015; Zuiderwijk et al., 2014):
F1. Search function using user input via keywords, filters, and data attributes
F2. Download dataset via link
F3. Register on portal as a new user
F4. Data can be cleaned, analyzed, enriched and linked
F5. Consult user path of data via extension diagrams about the use and link with other data
F6. Visualizing the results of data analyses on the basis of applied techniques including modeling and statistics
F7. Providing feedback to the data providers
F8. Uploading datasets by governmental organization stakeholders
F9. Adding metadata by registered data providers
F10. Following portal users
F11. Sharing findings via social media channels in terms of data-analyses and links to datasets
F12. Version management for uploaded datasets
F13. Data quality assessment occasion
F14. Offering tutorials regarding the use of open data
F15. Interaction mechanism between portal users about datasets or what can be learned from the use of open data
Each of the functions are derived from multiple publications, which are presented in table 1.

Table 1: Mentioned functions linked to publication authors

	author						
function	Zuiderwijk (2015)	Zuiderwijk et al. (2014)	Zuiderwijk et al. (2013)	Thorsby et al. (2016)	Colpaert et al. (2013)	Alexopoulos et al. (2014)	Iamamphai et al. (2016)
F1 Search	+	+	+	+	+		+
F2 Download	+	+	+			+	+
F3 Registering	+	+	+		+		
F4 Analysis	+	+	+	+		+	
F5 Pathways	+	+		+			
F6 Visualization	+	+	+			+	+
F7 Feedback	+	+	+		+	+	+
F8 Uploading	+		+		+	+	
F9 Metadata	+	+			+	+	
F10 Following users	+				+		
F11 Sharing conclusions	+				+		+
F12 Version management	+		+		+	+	
F13 Data quality	+	+	+			+	
F14 Tutorials	+	+		+			
F15 Interaction	+	+	+		+	+	+

From table 1 we learn that the earlier provides list of fifteen functions is acknowledged in several scientific publications. The functions represents available technology that can be used as means to apply pathways in practice. When designing the pathways, we assume these functions to enable open data analyses via step-by-step tasks to be carried out by the users.
According to European Data Portal (2017) the use of OGD can be improved by means of the following three suggestions for improvement: (i) Users of the portal should be able to interact with each other for exchanging feedback, this can be facilitated through an interaction mechanism on the open data portal (European Data Portal, 2017). (ii) Implementing a one-stop data shop, in which users select and download real-time data (European Data Portal, 2017). (iii) Digital transformation within public administration processes should be stimulated in the light of open data usage (European Data Portal, 2017).
In conclusion, there is a void in literature regarding the context of portal usage. This void can be filled by means of pathways that bridges the gap between functions and the use of open data for user

goals. The research method for achieving research results are described in the next section.

3. Research Method

From the previous section we have found a need to bridge available technology with the data analyses process carried out by users. The goal of this research is to develop pathways. For designing pathways we make use of the Information Systems Research Framework defined by Hevner (2004). This framework consists of several elements that are necessary to produce the productive application of information technology in form of pathways. The framework has three main aspects: environment, information systems research, and knowledge base. Hereby the environment represents the problem space with people, organizations and technology. Furthermore, the knowledge base consists of raw material to accomplish the information systems research. Therefore, the information from the environment and knowledge base are input for information systems research. The information systems research starts with the development of pathways because the use of open data lags behind among end-users. After the draw up of the pathways, they are evaluated by interviewing five experts for possible suggestions to improve the pathways. Consequently, after the assessment of pathways, the initial pathway designs are refined. The outline of this section is by means of three following subsections.

Subsection 3.1 explains the environment. Followed by the knowledge base elements in 3.2. Subsection 3.3 clarifies when, for input information in this research, is stepped in the environment for user needs or otherwise in the knowledge base for applicable knowledge. Finally, 3.4 presents the design science steps are explained as representation for the information systems research as part of the framework that is explained before.

3.1 Environment of problem space

The focus on the environment, which represents the problem space, is from perspective of the following stakeholders. First, the data providers, with the role of uploading OGD via a portal in reaction to data requests, as a source for trust and transparency from the government to civilians. They want to upload open data on an easy way. Secondly, the end-users for who the pathways are developed. They have a role as users of open data, but don't have all necessary capabilities and/or knowledge to fulfill this role efficiently. The latter has to do with the diversity of open data offerings, which cause an organizational problem for the portal owners. The government want trust and transparency, as well as economic development, but usage of open data lags behind. The available technology to analyze open data is not in balance with capabilities or experience of end-users. Programming or statistics require practice oriented skills that is time-consuming. But ready-to-use functions enable users to get most of the open data despite their lack of skills in programming for instance. The functions are an alternative to complex tools and can be developed in java which is an platform independent and object-oriented programming language with written libraries (Charatan & Kans, 2009). This makes it suitable to build-in techniques that are part of the pathways as an extension of a portal. It can solve the organizational problem on the portal by offering guidance to users that have interest in open data, despite their capabilities.

3.2 Knowledge base of unused materials

Furthermore, part of the knowledge base within this article, is distinguished between foundations and methodologies. Within this article, the foundations include portal functions, that can be considered as instruments for end-users to effectively make use of open data. Also mathematical models constructs can be used for time-series, classification, regression, and cluster modeling to get insights from open data. Finally, as the center of the framework is information systems research, the pathways are defined, evaluated, and consequently adjusted according to received feedback. This feedback originate from five respondents.

We use a sequence of steps as a procedure to define each pathway because this setting makes the process systematic. Therefore, the focus is on deriving insights from open data instead of correctly using available functions. Also it can be improved more efficiently due to the stepwise structure of pathways. Pathways provide structural approaches that deal with the use goals to get value from open data. Different pathways represent different strategies to substantiate the use of technology. Hereby, the data analysis research steps are consistent and goal-oriented and controlled manner towards insights that have not been seen before. The foundations are used as applicable knowledge to Information Systems Research. Furthermore, methodologies that are used as part of the knowledge base, are represented by data analysis techniques that are used to develop steps as part of the pathways.

3.3 Stepping into environment or knowledge base?

The information systems research framework is separated into environment, information systems research and knowledge base. The question hereby is when environment or knowledge base are used as input for the information systems research? If user needs need to become clear we step into the environment, but when applicable knowledge is required, the knowledge base can provide the input information (Hevner et al., 2004). This all depends on the information systems research. The environment is for defining needs in the problem space while the knowledge base is meant for finding applicable knowledge (Hevner et al., 2004). When the problem space is clear, the knowledge base provides solutions that fit in design science steps to represent the information systems research. This means that when all interests are clear the current state of technology is analyzed in terms of functions from the knowledge base. Consequently, the environment is used to find

user needs in order to develop pathways with applicable material from the knowledge base.

3.4 Information systems research with design science

Part of the Information systems research framework is the evaluation of five pathways by interviewing five experts in the use of open data with different expertise. The design is iterative and the pathways are refined according to the interview results. Before this is done, the pathways are designed by means of design science steps. Therefore, section content that is part of the article is linked to research activities that results into assessed and refined pathways. Hereby, Design Science Research Methodology (DSRM) Process Model steps are used, which are originally defined by Peffers et al. (2007, p. 43). Every step has a link with article content which are as follows:
In the first step the problem is described and motivated, which is a void in literature that can be filled by means of pathways. The pathways enable guidance for end-users to pursue specific open data user goals. In the second step the functions for processing data, including suggestions for improvements, are used as the objectives for a solution. This is in order to bridge the gap between available data and usage via pathway steps. The third step, is the design and development of the pathways (see section 4). In the fourth step, there is a demonstration of the pathways to understand how they work (see section 4). The fifth step, is for evaluation of the pathways among five interviewed experts (see section 5). Finally, the sixth step is for communication of the conclusions, as well as suggestions for future research (see section 6).

4. Developing pathways

In this section we present preferences that make open data offerings more user-friendly on the long term, in 4.1. An overview of the assumptions, differences and pathways is provided in 4.2. The pathways are initially defined in 4.3. There are preferences for how the open data can be offered in a user friendly manner. The pathways guide the use of available information technology in form of portal functions as well as build-in techniques that enable open data analysis online on a portal. Hereby, the goal is to answer research questions and/or hypothesis that have been defined at the beginning of a data analytics project.

4.1 Preferences for open data offerings

Each of the categories indicate preferences for how open data should be offered to standard end-users (Dawes et al., 2016):

- What we learn from the category formats, is that user-friendly open data should not require proprietary software to view the dataset. Furthermore, it is preferably offered in semantic linked open data formats and named with URI-entities (Uniform Resource Identifier). It means that open data is searchable through an uniform context and navigable via a graph query-language (SPARQL) in the datasets. Furthermore, the open data are framed by means linked open data framework RDF (Resource Description Framework).
- What we learn from the category metadata is that the data providers should be contactable about datasets for which they are responsible for. Furthermore, the structure of the dataset should be available and updated whenever possible. The keywords that represent the dataset should be representing its content.
- What we learn from the category access is that a dataset should be downloadable in different formats. This includes HTML-webpages, in combination with navigational information concepts such as a map.
- What we learn from the category quality is that the quality of datasets depends on how the data is homogenously structured and standardized via corresponding attributes. This means that unstructured pieces of text should be avoided, as well as other unusual information dataset. Furthermore, mistakes in a dataset should be avoided such as incomplete dataset, wrong information in dataset, spelling mistakes, and double information.

The data usage criteria explain an interdependency between the defined pathways because data is a dependency factor for all four pathways. It can depend on the data quality whether to start using data earlier in a procedure compared to less qualitative good data. The latter implies that pre-processing is required to prepare the data. But if for example the metadata is good and complete, techniques can be used in an earlier stage because you don't need to analyze the data descriptively for example. That is when the context is known in advance due to the metadata itself. Also the access to and formats of the data that enable the use of API's in order to automate repeatable processes for data analyses tasks.
How the open data is presented (format), described (metadata), structured (quality) and accessible is important within the pathways. This indicates the influence that offering of open data has on user goals. By developing pathways we can provide a guidance to use open data in independent manner to carry out the tasks. But with flexibility space for taking into account differences of open data offerings, that means that steps allow to adjust to how open data is offered. In this next subsection the pathways are explained in more detail.

4.2 Development of pathways

The four pathways are created by using literature about open data. Furthermore, the use of the Dutch OGD portal has been used as an inspiration objective for defining the pathways (data.overheid.nl).

The pathways are evaluated after demonstrating the pathways to five respondents in section 5.
In this subsection four pathways are defined that form the structured basis for open data analysis. The following overview of pathways indicates the focus and differences between the pathways.

A pathway is hereby a stepwise approach towards building a model and coming to valuable results that enable end-users to answer predefined research questions and/or hypotheses. The main objective is to build a representational model by means of data analysis. Hereby, this is not the right approach or strategy, it's a point of departure for developing specific models that represent real situations in practice. An overview is provided regarding different assumptions that are used to develop pathways:

1. The first assumption is that entrepreneurs want to make informative decisions, based on calculated scenario's, to pursue investment portfolios. Hereby classification models are of interest.
2. Second assumption is that researchers are interested in developing models that are based on in real-life constructs such as institutions. Hereby, regression models are preferred because they enable to model continuous variables that represent reality observations from open data.
3. Third assumption is that time-series data is used to describe data behavior over time for statistical insights with different states of a problem field over time.
4. Finally, descriptive statistics are used to inform the user of open data to assess what is possible with the dataset in an earlier stage. This also included the separation of initial datasets into several subsets for example.

An overview of the four pathways is provided in order to make sure that it becomes clear what the purposes are for users.

- The first pathway is focused on discovering structure from the open data. With structure we mean how the interrelatedness of variables are comprised. This can be researched through the use of regression models to make prediction models for continuous variables. This enables insight in how different factors influence each other and to distinguish between important and less important measures.
- The second pathway uses classification for insights to calculate different predefined scenario's. This means that the open data is classified by training a classifier with open data in order to classify an outcome or dependent variable. The outcome is based on a scenario that has a characteristic series of values for the chosen independent variables. This approach is suitable to analyze scenarios.
- The third pathway is focused on analyzing time-series data by means of choosing appropriate distributions to model the behavior at different time-ranges. This demands flexible modeling and interpreting the data behavior on several occasions. For example, to decide if data can be represented by exponential distributions, poison distribution, normal distributions etc. This demands also interpreting the context of the dataset. Normal distribution is more likely for a dataset about cars passing by a high-speed road during the year, instead of the civilians in a growing city that fits exponential distribution for example. Because it is no coincidence when increasingly more people start moving to a city for a jobs. This kind of modeling demands insight in distributions that can fit the behavior of the dataset.
- The fourth pathway is based on cluster analyses without supervising the model be means of assumptions. The strategy for clustering can be used to discover subgroups of data objects with the same characteristics. This is called a cluster. Of course choosing the dependent variables and independent variables are crucial, as well as how the clusters are defined. So based on which kind of variables and their threshold for associations with a defined cluster, when assigning values.

Differences between pathways are as follows: The first pathway is based on supervised regression models. Supervised hereby means that the model construct has parameters that are calculated by means open data analyses. The second pathway is unsupervised and more focused on how to calculate scenario's that represent a possible future situation. With classification modeling different scenarios can be calculated, for example by means of a decision tree's, which enable to make informed decisions. The third pathway is based on supervised modeling and focused on time-series analysis whereby statistical insights at different states of the model can lead to insight for improvement opportunities. Finally, the fourth pathway is unsupervised and hereby focused on explorative analysis by means of cluster analysis. Each cluster contains data objects that have a shared set of characteristics. In the next subsection the pathways are initiated.

4.3 Defining initial pathways

In this subsection the four pathways are initially defined as a point of departure for further development. But before they are defined the following two conditions for the development of pathways are taken into account (Geels & Schot, 2007):

- The first one is that pathways are not deterministic, which implies that the steps in pathways are and cannot be automated, and therefore different users can get different results with the same pathway.
- The second condition is that the pathways are ideal types, which cause a need for care in the application of pathways whereby considerable arguments need to be balanced regarding the choices made within a pathway.

These conditions are taken into account for the development of initial pathways and make clear that the pathways are not useful if a user want to use a pathway as a deterministic process which can be fully automated without random influences. Also the pathways are not usable when the user has no background in the problem space for which data analysis is used as an instrument to substantiate strategic decision-making.
The initially defined pathways are therefore defined as follows:

Pathway 1: structure discovery
Step 1: Define the research questions or define hypothesis that you want to confirm or reject.
Step 2: Acquire data from the open data portal. This can be either manually, by downloading the data(set). Or automatically, if you want it to be a repeatable process on a URL.
Step 3: Parse the data for quality assessment, regarding its purpose. If this is acceptable, continue, otherwise repeat the previous step with other data on the portal.
Step 4: Filter relevant subsets out-of-data, link to other (external) data, ranges, categories etc.
Step 5: Mining the data, starting with descriptive statistics. If this is acceptable and sufficient information, for example with linear regression lines, to model continuous variables. This is parametric as you take assumptions on the function of the data model that you want to make representational for the data(set) by means of parameters (James et al., 2013: 61). Otherwise the use of statistical inference, in term of non-parametric model, whereby you use a technique to discover the structure of the data. For example, by means of support vector machines (Berthold & Hand, 2007: 181). Of course both methods need to fit the model with the data. Different measures can be used for this, for example mean squared error (MSE) (James et al., 2013: 29). This measure can be used to measure the average squared difference between estimated values and true observations in the dataset.
Step 6: Present the results in terms of visualizations to grasp the results in a research report.
Step 7: Refine according to visualizations of relevant level-of-data for insights to answer research questions. If this acceptable answer the research questions and/or hypotheses, otherwise repeat previous step.
Step 8: Share the visualizations with other users on the portal and/or social media channels.

Pathway 2: classifier insight
Step 1: Exploring, searching for data(sets) on themes, filters, and data attributes.
Step 2: Defining hypotheses according to the data exploration to be confirmed or rejected.
Step 3: Selecting data analyzing methods. For example a prediction data model, in terms of a random forest, which is a technique that decorrelates decision trees (James et al., 2013). For the classification of data.
Step 4: Acquiring data from portal, either manually or automatically to be repeatable from an URL.
Step 5: Parse the data for quality assessment for the user goal or purpose. If acceptable, continue, otherwise repeat previous step.
Step 6: Filter relevant attributes and possibly link to other (external) data.
Step 7: Mining the data(sets) by making a model and test its performance. For example with neural network as the modelling technique (Berthold & Hand, 2007: 269). Furthermore, validating it by means of k-fold cross validation to choose the best configuration (Berthold & Hand, 2007: 58). So, separating the dataset into multiple sets and testing performance of model configurations on each different subset.
Step 8: Visualize the results, and present the classes that are relevant to answer the hypotheses.
Step 9: Refine the results, according to insights from the visualizations.
Step 10: Share the results with others for feedback.

Pathway 3: time series analysis
Step 1: Searching for datasets or API/URLs on the portal.
Step 2: Acquiring or downloading with an automated process or saving the data in a database. This should be a database with a predefined structure for time series data.
Step 3: Using unsupervised machine learning, in terms of cluster analyses, to discover subgroups with shared characteristics within the data (James et al., 2013: 386).
Step 4: Defining research question and/or hypotheses. According to the discovered clusters.
Step 5: Pre-processing the data, link it to other data sources for enriching the data, being able to do relevant research.
Step 6: Using the results for time series forecasting. There are multiple states, and techniques that can be used to model the behavior. The states are as follows (Ragsdale, 2008: 486):

- Stationary, without trends over time.
- Non-stationary, including upward and downward trends in the data.
- Seasonal, including patterns in both stationary and non-stationary time series data.

Step 7: Parse the data for quality assessment to answer research questions and/or hypotheses. If acceptable continue, otherwise repeat previous step.
Step 8: To make a good forecasting model on the data, the following techniques can be used (Ragsdale, 2008):

- Moving average for highlighting long-term trend over short-term random fluctuations.
- Weighted moving average, the same as previous technique, but in addition to that with weights assigned to the moving averages.
- Exponential smoothing for exponentially decreasing weights over time assigned with exponential functions, representing the data behavior over time.
- Holt-Winters' method for additive seasonal effects. This is a forecasting model which consist of a forecasting equation, together with three additive smoothing

equations. One for the level, for the trend, and a seasonal component.

Step 9: Parse data for quality assessment, if acceptable continue to next step. Otherwise repeat aforementioned step.

Step 10: Apply goodness of fit for assessing model forecasting performance. For example, by means of the following techniques (Ragsdale, 2008: 487):

- Median Absolute Deviation (MAD), to measure the average deviation of each measure relative to the mean of a quantitative dataset.
- Mean Absolute Percentage Error (MAPE), to measure accuracy of models in terms of percentages.
- Mean Squared Error (MSE), to measure the average squared difference between estimated values and true observations in the dataset.
- Root Mean Squared Error (RMSE), in order to compare the forecasting errors for different models, with the same dataset.

Step 11: Representing the results in terms of visualization.

Step 12: Refining the previous steps, according to insights in the visualizations. If acceptable, confirm research results by answering the research questions and/or confirming/rejecting hypotheses.

Step 13: Share the research results in form of a research report.

Pathway 4: cluster hypothesis

Step 1: Choose a dataset on the basis of a preferred theme of interest.

Step 2: Acquire the dataset manually by downloading it to the hard disk of a local computer.

Step 3: Filter relevant subsets and link to other data from external sources. Select ranges of values for different data attributes if necessary for insights.

Step 4: Apply cluster analysis on the dataset, for discovering groups with shared characteristics (James et al., 2013 :386).

Step 5: Define hypotheses for each clusters.

Step 6: Mining the data, possibly with statistical inference or descriptive statistics, depending on the hypotheses.

Step 7: Test the model performance by means of goodness of fit (Ragsdale, 2008). For example, to assess whether two samples are drawn from the same distribution that represents the dataset.

Step 8: Representing the results in form of visualizations for insights.

Step 9: Refining the analysis according to derived insights from visualizations until acceptable.

Step 10: Share insights in form of an online available research report.

To conclude this section four pathways are expected to be valuable to answer research questions and/or conforming/rejecting hypothesis. We assume a hypothesis to be an expected explanation for something which is measurable quantitatively after statistical analyses. In the next section we evaluate the four pathways by interviewing five experts and applying received feedback.

5. Evaluating the pathways

In this section the four pathways are evaluated by interviewing five experts about the pathways and the use of open data with differing expertise. Each respondent has the following expertise:

1. Producing models and testing them consequently by analyzing open data to answer research questions and hypotheses.
2. Using open data for scientific analyses, for example to discover trends in traffic for logistical insights.
3. Open data is a source of trust and transparency for the government as well as a source for business opportunities on new locations wherefrom society can profit.
4. Open data enables new and more advantages by means of using open data.
5. Developing framework that merges different data sources in order to make big data analyses easier.

Subsection 5.1 presents the answers regarding the importance of each pathway on a scale from 1 to 5. In 5.2 the answers of the respondents are described after asking them if there are any missing steps and/or incorrect ordering of steps on the defined pathways. Consequently, in subsection 5.3 the suggestions for improvement of the pathways are described. Finally, in subsection 5.4 the pathways are adjusted by implementing the suggestions from previous subsection.

5.1 Evaluation of pathways

Table 2 indicates the presumed importance of pathways from perspective of each expert after presenting the four pathways to them. The score is measured on a scale from 1 to 5.

Table 2: Importance of pathway, Scale: 1= not very important, 5= Very important

Path way	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Aver age
1	3	4	5	5	5	4,4
2	4	5	5	5	5	4,8
3	4	5	5	5	4	4,6
4	4	5	5	5	4	4,6

According to the results in table 3 we can conclude that all pathways have a importance with a score of above 4 from all experts. This implies that the pathways are potentially valuable as guidance for end-users. In the next section we evaluate the pathways with more questions to the respondents.

5.2 Missing steps and ordering in pathways

Five respondents have been interviewed and asked if there were missing steps in the pathways. The results are present in table 3.

Table 3: Missing steps in the pathway?, Legend: Y=Yes (0 point) and N=No (1 point)

Path way	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Score
1	Y	Y	Y	Y	Y	0
2	Y	N	Y	Y	Y	1
3	Y	Y	N	N	Y	2
4	Y	N	Y	N	Y	2

From table 3 we can conclude that expert 1 and 5 indicate missing steps in all defined pathways. This implies that the experts prefer more specific pathways regarding the use of open data for scientific analyses. Furthermore, the other three experts are less critical in terms of missing steps, but they do have indicated them. Therefore, we can conclude that there is space for improvement of all four pathways. In table 4 the question to the expert is if the steps that form the different pathways have the right ordering. The answers to this question are presented in table 4.

Table 4: Right ordering of pathway steps?, Legend: Y=Yes (1 point), N=No (0 point)

Path way	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Score
1	Y	Y	Y	Y	Y	5
2	N	N	Y	Y	Y	3
3	N	Y	Y	Y	Y	4
4	N	N	Y	Y	Y	3

After viewing table 4 it becomes clear that the first pathway scores best among the five interviewed experts. In general the experts are of the opinion that the ordering of steps is right. But there are experts that have different views on the ordering of steps regarding pathway 2 until 4. Therefore there is also space for improving the pathways in terms of ordering. In the next subsection the more detailed feedback from experts is presented before its implementation in revised versions of the different pathways.

5.3 Feedback for improving pathways

The following feedback is provided by the respondents. Open data is considered to be a source for answering research questions and testing hypotheses. Therefore, developing models and testing them statistically for discovering trends can lead to insights. For example, for business opportunities that lead to new economic venues, and hereby benefitting society. Also OGD enables society to demand better policies, as it is also a source for trust and transparency from perspective of the government.

According to the experts all the pathways need a feedback loop to scrutinize the data quality, and sharing this process with others is even more desirable. This is in pathway 1, between step 6 and 8, in pathway 2 between step 8 and 9, and in pathway 4. *"Data quality is the most challenging when using open data".* Furthermore, you first start with the research questions, and then collect the data (suggestion for pathway 2 & 3). Preferably you answer the research question and/or confirm/reject the hypotheses on the last step (suggestion for pathway 1). *"Incomplete dataset, structure of dataset, and non-machine readable datasets are mostly challenging".* Steps 3, 4 and 5 depends on how big the dataset is. If there is a big dataset, it needs to be separated into small subsets, and in case of small datasets it needs to be updated more frequently. Between step 3 and 4 of pathway 3, as well as step 5 and 7 of pathway 3 need more explanation and in general pathway 3 contains too much information. Because it depends on practical terms what to extract from the data. *"Civil service behavior, by providing OGD, and getting data users to demand and use open data is important".*

Regarding the last step of pathway 1, the publication in academic journal is part of the procedure. Alongside setting up data analysis experiments for different regions and domains. Finally, there is need for guidelines regarding security and privacy issues in terms of policy after a risk assessment of disclosing OGD. *"All methods depend on objectives at beginning of pathway".* Regarding pathway 3, the evaluation method depends on the applied techniques and objectives at the beginning of the pathway.

In the next subsection the pathways are improved with the received feedback.

5.4 Updating pathways after feedback

The feedback from the respondents is integrated in the pathways.

In subsection 5.1 we evaluated the pathways among five experts in terms of importance (1-5). In 5.2 we evaluated the pathways in terms of possible missing steps (yes/no) and the right ordering of steps (yes/no). Consequently, in 5.3 the feedback on pathways is discussed. In this subsection 5.4 the feedback is processed by improving the pathways accordingly. Remarkable is that the

interviewed experts are mainly more focused on data quality and how to interact with the data owner. The pathways are therefore adjusted on the basis of this and got feedback loops with the data providers. The adjusted pathways are as follows:

Pathway 1: structure discovery

Step 1: Define the research questions and/or hypotheses that you want to confirm or reject.

Step 2: Acquire data from the open data portal. This can be either manually, by downloading the data(set). Or automatically, if you want it to be a repeatable process on an URL.

Extra step A: The acquired data has influence on how the following steps: 3, 4 , and 5 are carried out and/or defined. If there is a big dataset, it needs to be separated into subsets. Otherwise data is analyzed inefficiently, which is a waste of time. If there is only a small dataset, it can be either updated by the data provider after feedback to the data provider. Or link to other possible datasets.

Step 3: Parse the data for quality assessment, regarding its purpose. If this is acceptable, continue, otherwise repeat the previous step with other data on the portal.

Extra step B: Provide feedback to data provider for scrutinizing the data quality.

Extra step C: Share data quality issues with others similar dataset users to enable interaction about the data quality.

Step 4: Filter relevant subsets out-of-data, and link to other (external) data, ranges, categories etc.

Step 5: Mining the data, starting with descriptive statistics. If this is acceptable and sufficient information. For example with linear regression lines for the right variables. This is parametric as you take assumptions on the function of the data model that you want to make representational for the data(set) by means of parameters (James et al., 2013: 61). Otherwise the use of statistical inference, in term of non-parametric model, whereby you use a technique to discover the structure of the data. For example with Support Vector Machines (SVM) (Berthold & Hand, 2007: 181). Of course both methods need to fit the model with the data. Different measures can be used for this, for example Mean Squared Error (MSE) (James et al., 2013: 29). This measure can be used to measure the average squared difference between estimated values and true observations in the dataset.

Step 6: Present the results in terms for visualizations that grasp the results in a research report.

Extra step D: Provide feedback to data providers, for scrutinizing the data quality.

Step 7: Refine according to visualizations the relevant level-of-data for insights to answer research questions. If this acceptable answer the research questions and/or hypotheses, otherwise repeat previous step.

Step 8: Share the visualizations with other users on the portal and/or social media channels.

Extra step E: Answer the research questions and hypotheses according to derived results.

Pathway 2: classifier insight

Extra step A: Start with defining research questions.

Step 1: Exploring, searching for data(sets) on themes, filters, and data attributes.

Extra step B: Provide feedback to the data providers if something is missing or wrong with the data.

Step 2: Defining hypotheses according to the data exploration to be confirmed or rejected.

Step 3: Selecting data analyzing methods. For example a prediction data model, in terms of a random forest. This is a technique that decorrelates decision trees for the classification of data (James et al., 2013).

Extra step C: Explain what kind of information you want, how you would do this in order to understand which additional data you need.

Step 4: Acquiring data for portal, either manually or automatically to be repeatable from an URL.

Step 5: Parse the data for quality assessment for its user goal or purpose. If acceptable, continue, otherwise repeat previous step.

Extra explanation for step 5: The data quality criteria should be defined in order to decide if the data is useful for answering the research questions and/or confirming or rejecting the predefined hypotheses.

Step 6: Filter relevant attributes and possibly link to other (external) data.

Step 7: Mining the data(sets), make model and test its performance. For example with neural network as the modelling technique (Berthold & Hand, 2007: 269). Furthermore, validating it by means of k-fold cross validation to choose the best configuration (Berthold & Hand, 2007: 58). So, separating the dataset into multiple sets and testing performance of model configurations on each different subset.

Step 8: Visualize the results, and present the classes that are relevant to answer the hypotheses.

Step 9: Refine, according to insights from the visualizations.

Extra step D: Provide feedback to data providers regarding the quality of used open data.

Step 10: Share the results with others for feedback.

Pathway 3: time series analysis

Extra step A: Start defining the research questions.

Step 1: Searching for datasets or API/URLs on the Dutch OGD portal.

Step 2: Acquiring or downloading with an automated process or saving the data in a database. This should be a database with a predefined structure for time series data.

Step 3: Using unsupervised machine learning, in terms of cluster analyses, to discover sub groups with shared characteristics within the data (James et al., 2013: 386).

Extra step B: Interpret the results in the light of relevant further research.

Step 4: Defining research questions and/or hypotheses. According to the discovered clusters.

Extra step C: Translate the research questions and hypotheses into a data model that can answer all the questions.
Step 5: Pre-processing the data, link it to other data sources for enriching the data, being able to do relevant research.
Step 6: Using the results for time series forecasting. There are multiple states, and techniques that can be used to model the behavior. The states are as follows (Ragsdale, 2008: 486):

- Stationary , without trends over time.
- Non-stationary, including upward and downward trends in the data.
- Seasonal, including patterns in both stationary and non-stationary time series data.

Step 7: Parse the data for quality assessment to answer research questions and/or hypotheses. If acceptable continue, otherwise repeat previous step.
Extra step D: Define the criteria that you want to use for data quality assessment, this depends on what kind of research questions you would like to answer and the hypotheses that you want to reject or confirm. Then continue to the quality assessment to improve it by providing feedback to data providers.
Step 8: To make a good forecasting model on the data, the following techniques can be used (Ragsdale, 2008):

- Moving average for highlighting long-term trend over short-term random fluctuations.
- Weighted moving average, same as previous techniques but assigned weights to the moving averages.
- Exponential smoothing for exponentially decreasing weights over time assigned with exponential functions, representing the data behavior over time.
- Holt Winter's method for additive seasonal effects. This is a forecasting model which consist of a forecasting equation, together with three additive smoothing equations. One for the level, one for the trend, and a seasonal component.

Step 9: Parse the data for quality assessment, if acceptable continue. Otherwise repeat aforementioned step.
Step 10: Apply goodness of fit for assessing model forecasting performance. For example, by means of the following techniques (Ragsdale, 2008: 487):

- Median Absolute Deviation (MAD), to measure the average deviation of each measure relative to the mean of a quantitative dataset.
- Mean Absolute Percentage Error (MAPE), to measure accuracy of models in terms of percentages.
- Mean Squared Error (MSE), to measure the average squared difference between estimated values and true observations in the data.
- Root Mean Squared Error (RMSE), in order to compare the forecasting errors for different models, with the same dataset.

Extra step E: The chosen evaluation technique depends on which model is chosen for data analyses results.
Step 11: Representing the results in terms of visualization.
Step 12: Refining the previous steps, according to insights in the visualizations. If acceptable, confirm research results by answering the research questions and/or confirming/rejecting hypotheses.
Step 13: Share the research results in form of a research report.

Pathway 4: cluster hypotheses
Extra step A: Do quality assessment on potential data.
Step 1: Choose a dataset on the basis of a preferred theme of interest.
Step 2: Acquire the dataset manually by downloading it to the hard disk of a personal computer.
Extra step B: Provide feedback to data providers regarding the chosen data.
Extra step C: Share data quality assessment with other users for feedback. In order to combine it with other data to improve that data quality. Hereby it is important that links between data are recognized and added where possible, also if an identifier column is missing for instance.
Step 3: Filter relevant data subsets, and link to other data from external sources. Select ranges of values for different data attributes if necessary.
Step 4: Apply cluster analysis on the dataset, for discovering groups with shared characteristics (James et al., 2013 :386).
Step 5: Define hypotheses according to defined clusters.
Step 6: Mining the data, possibly with statistical inference or descriptive statistics, depending on the hypotheses.
Step 7: Test the model performance by means of goodness of fit (Ragsdale, 2008). This measure can be used for statistical hypothesis testing. To assess for example whether two samples are drawn from the same distribution that represents the dataset.
Step 8: Representing the results in form of visualizations for insights.
Step 9: Refining the analysis according to derived insights from visualizations until acceptable.
Step 10: Share insights in form of an online available research report.

After feedback, it still does not mean that the pathways are perfect. Only five experts are interviewed for feedback although the pathways are developed for more end-users in general. More feedback might by retrieved if more experts had been interviewed. As we have come to the end of the research results, in section 6 the conclusion and possible future research are presented.

6. Conclusions and future research

Value creation with OGD is challenging and in this paper we explored the concept of pathways. They are designed to support value creation with open data. Pathways enable to structure the approach for achieving user goals. This can save search time for data users because they can make use of guidance for their data-

analysis projects. Although approaches for analyzing open data can vary per user, the pathways are a tool that can be chosen by users that see value in it. It is not a requirement for using open data but a guidance tool to save time and making the use of open data more user-friendly. The user-friendliness of pathways based on the bridge between available data analyses, technology/tools, and decisions by data users. This research has been a first step in doing this, because current research has mainly being focused on data analyses itself.

The complexity about the development of pathways is to structure them in such a way that it leads to valuable data analyses, while taking into account the usage of available technology. Hereby several functions that can be part of an OGD portal were derived from literature as an inspiration. The systematic research with clear results that form the basis of the scientific contribution. This have led us to future research opportunities which are defined below in terms of indicative research questions.

The following research questions help to develop the pathways further in future research efforts:

1. *Which approach is useful for optimizing queries to pathways, by means of representative neighboring keywords?*
2. *Can data users be enabled to add pathways to the existing collection on a user-friendly manner?*
3. *Is it hereby possible to enable end users to adjust pathways according to their perceptions on it, and store the changes on pathways on an individual basis?*

The first research question is according to section 2 in which functions are presented. The question is here is how pathways would fit in a portal, that is surrounded by other functions. More research is needed regarding the future implementation and use of pathways by end users.

The second research question is according to section 4 where the first pathways are developed. It would be interesting to enable end users to construct their own pathways and add them to the existing collection.

The third research question is based on section 5, in which the initially defined pathways are adjusted according to received feedback from respondents. If end users can be enabled to adjust the pathway in use, it would be beneficial for future users with similar needs. Therefore, it is interesting to do future research about storing individual changes to pathways, according to different perceptions, on a specially designed database.

7. REFERENCES

[1] Alexopoulos, C., Loukis, E. and Charalabidis, Y. (2014). A Platform for Closing the Open Data Feedback Loop based on web2.0 functionality. *JeDem eJournal Of Democracy & Open Government*, Vol. 6, Iss. 1, pp. 62-68.

[2] Berthold, M. & Hand, D.J. (2007). Intelligent Data Analysis, An Introduction. *2nd revised and extended edition*, Springer.

[3] Charatan, Q. and Kans, A. (2009). Java in Two Semesters, *3th edition*. McGraw-Hill Education, Berkshire.

[4] Chen, H., Chiang, R.H.L. and Storey, V.C. (2012). Business Intelligence and Analytics: From Big Data to big Impact. *MIS Quarterly*, Vol. 36, No. 4, pp. 1165-1188.

[5] Colpaert, P., Joye, S., Mechant, P., Mannens, E. and Van de Walle, R. (2013). The 5 Stars of Open Data Portals. *Proceedings MeTTeG*, Vol. 7, pp. 61-67. http://pieter.pm/5stardataportals/

[6] Dawes, S.D., Vidiasova, L., and Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, Vol. 33 Iss. 1 pp. 15-27. https://doi.org/10.1016/j.giq.2016.01.003.

[7] De Rosnay, M.D. and Janssen, K. (2014). Legal and Institutional Challenges for Opening Data across Public Sectors: Towards Common Policy Solutions. *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 9, Iss. 3. Doi:10.4067/S0718-1876201400300002.

[8] European Data Portal. (2017). *Open data maturity in Europe 2017. Open data for a European data community*. Available on: https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf

[9] Geels, F.W. and Schot, J. (2007). Typology of sociotechnical transition pathways. *Research Policy*, Vol. 36, Iss. 3, pp. 399-417. https://doi.org/10.1016/j.respol.2007.01.003

[10] Hevner, A.R., March, S.T., Park, J., and Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, Vol. 28 No. 1, pp. 75-105.

[11] Huijboom, N. and Van den Broek, T. (2011). Open Data: an International comparison of strategies. *European Journal of ePractise*, No. 12. Available on: http://unpan1.un.org/intradoc/groups/public/documents/UN-DPADM/UNPAN046727.pdf

[12] Iamamphai, P., Noymance, J., San -Um, W. and Pasupa, K. (2016). Investigations and Comparisons of Government Open Data Websites through Systematic Functional Analysis and Efficient Promotion Approach. *The 2016 Management and Innovation Technology International Conference (MITiCON-2016)*, pp. 142-147.

[13] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning, with applications in R. *Springer*. Doi: 10.1007/978-1-4614-7138-7. http://www-bcf.usc.edu/~gareth/ISL/

[14] Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, Vol. 29 pp. 258-268. https://doi.org/10.1080/10580530.2012.716740.

[15] Nugroho, R.P. (2013). A comparison of open data policies in different countries. *Master thesis*, Delft University of Technology.

[16] Oxford. (2018). Pathway. English Oxford Living Dictionaries. Retrieved on 18-09-2018 from: https://en.oxforddictionaries.com/definition/pathway

[17] Peffers, K., Tuunanen, T., Rothenberger, M.A., and Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, Vol. 24 Iss. 3 pp. 45-78.

[18] Ragsdale, C.T. (2008). Spreadsheet Modeling & Decision Analysis. *Fifth edition*

[19] Thorsby, J., Stowers, G.N.L., Wolslegel, K. and Tumbuan, E. (2016). Understanding the content and features of open data portals in American cities. *Government Information Quarterly*. http://dx.doi.org/10.1016/j.giq.2016.07.001.

[20] Ubaldi, B. (2013). Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Working Papers on Public Governance*, No. 22, OECD publishing, Paris. http://dx.doi.org/10.1787/sk46bj4r03s7-en.

[21] Zuiderwijk, A. (2015). Open Data Infrastructures, The design of an infrastructure to enhance the coordination of open data use. *PhD thesis*.

[22] Zuiderwijk, A., Helbig, N., Gil-García, J.R. and Janssen, M. (2014a). Special Issue on Innovation through Open Data – A Review of the state-of-the-Art and an Emerging Research Agenda: Guest Editors' Introduction. *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 9, Iss. 2, pp. 1-13. Doi: 10.4067/S0718-18762014000200001.

[23] Zuiderwijk, A., Janssen, M. & Parnia, A. (2013). The complementarity of open data infrastructures: An analysis of functionalities. *Paper presented at the 14th Annual International Conference on Digital Government Research*, Quebec, Canada.

[24] Zuiderwijk, A., Janssen, M. and Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. *Information Polity*, Vol. 19, IOS Press, pp. 17-33. DOI: 10.3233/IP-140329.