DEVELOPING AN OPEN DATA LIFECYLCE MODEL BASED ON LITERATURE AND PRACTICE

Veenstra, Anne Fleur van, TNO, Brassersplein 2, 2612 CT Delft, The Netherlands, annefleur.vanveenstra@tno.nl

Broek, Tijs van den, TNO, Brassersplein 2, 2612 CT Delft, The Netherlands, tijs.vandenbroek@tno.nl

Bastiaans, Mark, TNO, Brassersplein 2, 2612 CT Delft, The Netherlands, mark.bastiaans@tno.nl

Plas, Arjanna van der, TNO, Brassersplein 2, 2612 CT Delft, The Netherlands, arjanna.vanderplas@tno.nl

Abstract

Government organizations around the world have developed open data strategies to increase transparency and enable re-use of their data. Semi-public organizations follow, aiming to use open data also for commercial purposes. However, in practice, many organizations find the process of opening up their data cumbersome and they do not know which steps to take. Lifecycle models can guide the process of opening up data. Therefore, this paper develops an open data lifecycle model based on literature and practice. First, using existing open data lifecycle models this paper identifies generic phases of opening up data. Then, investigating the process of opening up data in a semi-public organization in the Netherlands, the lifecycle model is refined and detailed. While existing open data lifecycle models focus mainly on technical aspects of opening up data, our case study shows that involving stakeholders within and outside the organization – also from an early stage – is important for the success of open data. This spurs re-use and allows for open data to be embedded into the organizational strategy and work processes. The lifecycle model developed in this paper thus allows for the development of an open data strategy, rather than merely opening up individual datasets.

Keywords: Open Data, Open Government, Lifecycle Model, Case Study.

1 Introduction

Open data gained momentum since President Obama of the United States announced his 'open government' strategy (McDermott, 2010). Since then, governments around the world have adopted 'openness as a strategy' for their organizations to become more transparent and thereby accountable to citizens (Jaeger & Bertot, 2010). Furthermore, open data is increasingly seen as a strategy to realize economic activity (Harrison & Pardo, 2012) by enabling re-use of data. By now, also semi-public organizations, such as cultural heritage foundations, public transport organizations, network operators and research institutes, have adopted open data strategies. For these organizations the purpose of opening up their data extends beyond increasing transparency and accountability or increasing the economic value of their data for others to re-use. They are also looking for ways to enhance the value of their data for their own (commercial) purpose.

However, organizations often find the process of opening data cumbersome (Janssen, Charalabidis & Zuiderwijk, 2012). They are often unaware which steps to take in the process of opening up data. Lifecycle models are used to guide the development of open data, see, for instance, Alani et al., 2007; Curtin, 2010; Hausenblas, 2011; Hyland & Wood, 2011 and Janssen & Zuiderwijk, 2012. However, most of these models have a strong technical orientation, while also organizational challenges to open data exist. Furthermore, few of these models have been developed based on empirical investigations. Therefore, this paper develops an open data lifecycle model that is based on literature and practice.

The development of this open data lifecycle model takes place in two steps. First, existing open data models are compared to identify generic phases that organizations opening up data go through. Then, based on a case study of a research and technology organization (RTO) in the Netherlands the model is validated and detailed, including the specific activities and roles to adopt in every phase. The next section presents existing lifecycle models and compares them, formulating five generic phases that all organizations go through to open up their data. The third section describes the case study of an RTO in the Netherlands. The fourth section presents the main findings from the case study by formulating the refined open data lifecycle model. Section five discusses these findings and, finally, section six formulates conclusions and recommendations for further research.

2 Lifecycle models of open data

While many government organizations aim to open up their data, the process of opening data is usually cumbersome and many challenges persist (see, for instance, Janssen et al., 2012). Organizations are struggling with opening up their data and how to address the barriers that exist. One way of structurally capturing challenges and addressing them is by formulating a lifecycle model. A lifecycle is an examination of a system or proposed system that addresses all phases of its existence (Blanchard & Fabrycky, 2006). Often lifecycle models are associated with the development of tangible products, services or assets, such as software development (Stallinger et al., 2011). In that context, a lifecycle model defines the processes that apply to software throughout its lifecycle. Alongside these processes, it also defines activities, tasks and outcomes for every phase of the lifecycle and serves as a common body of language.

The purpose of lifecycle models is twofold: they capture the development of certain phenomena (describing) and predict the next steps in the development (prescribing) (Lane & Richardson, 2011). In e-government, lifecycle models help researchers to describe the process to an e-government initiative, instead of the outcome (see, for example, Tsai, Choi & Perry, 2009). In contrast to maturity models (see, for example, Kalampokis, Tambouris & Tarabanis (2011) for an open data maturity model), lifecycle models do not prescribe organizational stages of the software development process. Still, the process steps of developing information systems differ among situations. Hence, normative lifecycle models are often criticized as non-situational (White Baker, 2010).

Also in literature on open data, a variety of lifecycle models can be found describing the process of opening up data and guiding organizations through this process. Based on an extensive literature search, we found eight open data lifecycle models: Alani et al. (2007), Curtin (2010), Ferrara et al. 2011), Hausenblas (2011); Hyland (2010), Hyland & Wood (2011), Villazon-Terrazas et al. (2011), and Janssen & Zuiderwijk (2012). The number of lifecycle models urges for comparison and synthesis (Ruparelia, 2010). Therefore, this paper compares existing models of open data in order to develop a lifecycle model that includes all relevant dimensions of open data and that can be validated in practice. Table 1 identifies the phases and activities in the lifecycle models that were found in literature. The column on the right lists the subsequent steps formulated in these models. Then, shown in the middle column of table 1, we formulated common actions identified based on these existing models. Finally, we identified five common phases of opening up data: identification, preparation, publication, re-use and evaluation. These are shown in the left-most column of table 1.

Lifecycle phase	**Steps per phase**	**Activities in literature**
Identification	Setting the strategy	- Setting aims of open data (Alani et al., 2007) - Data awareness (Hausenblas, 2011) - Deciding on making data available (Janssen & Zuiderwijk, 2012)
	Selecting the data	- Collecting databases (Alani et al., 2007) - Approving the open datasets (Curtin, 2010) - Supporting the data selection (Ferrara et al., 2011) - Finding data for potential re-use (Hyland, 2010) - Obtaining a copy of the models of the databases (Hyland & Wood, 2011) - Obtaining data extracts or create replicable data (Hyland & Wood, 2011) - Identifying real life objects in data (Hyland & Wood, 2011) - Identifying data (Janssen & Zuiderwijk, 2012)
Preparation	Setting requirements	- Analysing requirements (Alani et al., 2007)
	Modelling and describing data	- Describing data and give it context (Hyland, 2010; Hyland & Wood, 2011) - Specifying, defining and analysing the data (Villazon-Terrazas et al., 2011) - Design and build an ontology for the data (Alani et al., 2007; Ferrara et al., 2011; Hausenblas, 2011; Hyland & Wood, 2011; Villazon-Terrazas et al., 2011) - Defining a schema pattern for the Unique Resource Identifier (Ferrara et al., 2011; Hyland, 2010; Hyland & Wood, 2011) - Planning for persistence of data, e.g., Persistent Uniform Resource Locators (Hyland, 2010)
	Converting to machine-readable data format	- Generating the data (Villazon-Terrazas et al., 2011) - Convert the data to machine-readable format (Alani et al. 2007; Ferrara et al., 2011; Hyland & Wood, 2011; Villazon-Terrazas et al., 2011) - Cleaning the data (Villazon-Terrazas et al., 2011)
	Linking data	- Mapping the data and ontology to existing ontologies and database (Alani et al. 2007) - Linking data to existing data (Villazon-Terrazas et al., 2011)
	Storing data	- Storing data in a datastore (Ferrara et al., 2011)

Publication	Publication of data	- Publishing data (Curtin, 2010; Hausenblas, 2011; Hyland, 2010; Hyland & Wood, 2011; Janssen & Zuiderwijk, 2012; Villazon-Terrazas et al., 2011)
	Publication of metadata	- Attaching data provenance for tracking (Curtin, 2010) - Publishing metadata (Villazon-Terrazas et al., 2011)
Re-use	Exploiting of published data	- Creating an online data catalogue of published data for data discovery (Hausenblas, 2011; Hyland, 2010; Janssen & Zuiderwijk, 2012; Villazon-terrazas et al. 2011) - Managing access rights to the dataset (Ferrara et al., 2011) - Exploiting the data (Villazon-Terrazas et al., 2011)
	Data management	- Maintaining of data (Hyland & Wood, 2011) - Processing and visualizing the data (Janssen & Zuiderwijk, 2012) - Discussing the quality and relevance of the data (Janssen & Zuiderwijk, 2012) - Recommending existing and future data (Janssen & Zuiderwijk, 2012)
Evaluation	Developing business propositions	- Developing use cases of data (Hausenblas, 2011)
	Monitoring and improving data	- Monitoring data re-use (Janssen & Zuiderwijk, 2012) - Integrating and improving data (Hausenblas, 2011; Janssen & Zuiderwijk, 2012)

Table 1. Open data lifecycle phases and the actions that are undertaken in every phase.

However, few of these models are based on practical experience, or have been validated in practice. Furthermore, these models have all been based on cases of open data in the public sector, focusing strongly on merely making sure that data are opened up to the public (following the notion of 'compliance' with open data) rather than ensuring that open data becomes part of the strategic mission of the organization. Therefore, we found that there is a need to validate an open data lifecycle using a case study of a semi-public organization aiming to embed open data in its strategy and work processes.

3 Case study

3.1 Case study methodology

To investigate the process of opening data we use an interpretivist methodology for in-depth research of a single organizational case study, fitting its complexity (e.g. Klein & Myers, 1999). Interpretivist research is "aimed at producing an understanding of the context of the information system, and the process whereby the information system influences and is influenced by the context" (Walsham, 2006). In the previous section, the different phases of the lifecycle model and the steps to be undertaken in these phases were identified. Using a longitudinal case study approach we aim to validate and refine the subsequent phases of the lifecycle model. The case selected is TNO (www.tno.nl), the national RTO of the Netherlands. This case was selected as the organization is in the middle of opening up its data to the public. This means that data could be collected during the implementation of the open data strategy.

For analysing the case study we use a triangulation of methods (Mingers, 2001), including action research and semi-structured interviews. This combination of methods aimed to both capture the variety of the actions that were undertaken, and capture the involvement and attitude of stakeholders and evaluate the actions at the same time. The stakeholders were sampled based on their involvement

in the process of opening up data as well as on their role in the organization. The action research consisted of the research team keeping track of actions that were undertaken throughout the process of opening up data, which started in September 2012 and continued until February 2013.

Subsequently, we validated these findings by conducting eight semi-structured interviews. These interviews were held with five data owners, a director or research, a strategist and an information manager who were all invited to reflect on the process of opening up data and on their role in this process. The interviews were held in November 2012 and January 2013 and lasted 45 minutes on average. Table 2 provides an overview of the interviewees. Central questions concerned the strategic choices for opening up data of the RTO, their experiences with opening data, the actions that were undertaken and their significance, as well as the involvement of significant stakeholders. Based on the findings the steps to be taken during the process of opening up data were identified.

Function	**Role in the open data process**
Director of research	Top management
Strategic advisor	Top management
Information manager	Information Manager
Senior research scientist ‘Employment data’	Data owner
Researcher / consultant ‘Employment data’	Data owner
Software engineer ‘Geological data’	Data owner
Senior research scientist ‘Traffic data’	Data owner
Junior research scientist ‘Traffic data’	Data owner

Table 2. Overview of the interviewees and their function.

3.2 Opening up data in an RTO

TNO is the national RTO of the Netherlands and can thus be considered a semi-public organization. The organization has long opened some of its research data to the public; for some time, the organization even was the largest contributor of datasets to the national open data portal data.overheid.nl. However, opening data was not undertaken in a structural manner, but took place incidentally. The decision of Ministry of the Interior to build an open data portal, also spurred the attention for open data within the RTO that began to realize that it may have a responsibility to open up its data. Therefore, a first meeting was organized, bringing together those stakeholders in the organization that have an interest in opening up their data.

During this meeting the purpose of opening data by the RTO was identified to consist of three different reasons. Firstly, opening up data is seen as a necessity for transparency, for example to show how research data are gathered and how they are structured. Secondly, the data of the RTO can be re-used by others to develop new services and stimulate economic development. This is especially relevant as many research projects of the RTO are funded by the government and these data can thus be seen as a public good. Thirdly, the RTO also has a commercial interest in open data. Therefore, the RTO is looking for ways to use their data to develop new commercial activities, for example by forging strategic partnerships with other data owning organizations.

To develop a structural way of opening data, during the fall of 2012 the RTO undertook a pilot project in which a few datasets were opened to the public. During this pilot project a few steps were undertaken. Firstly, suitable datasets that could be opened up were identified and the data owners of

these datasets were invited to participate in this pilot. Two datasets were identified and subsequently prepared for opening up. Secondly, the datasets were opened up especially to take part in a *hackathon*, a one-day workshop in which 150 participants could use the data to develop their own services. The hackathon was organized by the city of Rotterdam in October 2012 and aimed to promote the commercial use of public data in an urban environment. Data owners provided and pitched their data to teams of voluntary programmers. Several prizes (ranging from 500 to 3000 euro) were granted to the winning teams to stimulate the development of apps in specific areas of re-use: healthcare, business, tourism and mobility. And thirdly, these activities were evaluated with the data owners and other stakeholders that were involved. Based on these activities, a refined open data lifecycle was developed capturing lessons learned.

4 A refined open data lifecycle model

4.1 Identification

The first phase of opening data comprised the definition of the process of opening up data and the identification of data that were to be opened. In the case of the RTO, a meeting was organized in which all relevant organizational stakeholders were involved. Furthermore, as the purpose of the pilot project was to open up data during a hackathon, contact was made with the hacking community to identify which data would be interesting for re-use. We found this phase to consisted of two steps: strategy setting and identification of data for opening up.

4.1.1 Setting the strategy

In the case of the RTO, the strategy setting step of the pilot project was limited. It consisted of top management deciding to undertake the pilot project to investigate how the process of opening up data works in practice. The roles that were involved in this step included the information manager and top management, as well as the community manager. The former two roles were necessary to define the scope of the project internally, while the latter role connected with potential users. Top management support was found to be of critical importance – even though it was only a pilot project. As there were no procedures yet that could be followed when opening up data, this support was necessary as it meant that risks could be mitigated. This was especially relevant as a result of the different types of interest the RTO has in relation to open data. The director: *"Our role in the world of open data is quite interesting: we aim for transparency, but at the same time we have to make money. Open data may be even more challenging for us than for public organizations."*

4.1.2 Selecting the data

During the second step of the identification phase, the information manager and the data owners identified datasets to be opened up during the hackathon. To identify which data could be opened up, we used ePSI Platform (2013) guidelines. These state that datasets can only be opened up when they comply to regulations regarding ownership (including intellectual property), privacy and security. Using a longlist of available datasets that comply to the above-mentioned criteria, the most meaningful datasets were selected: the shortlist. In this pilot project, this meant that three datasets were selected from different domains: health, transport and geology. We found that in this step the support of the data owners was of great importance as they need to be willing to prepare the data. Not all data owners are familiar with open data yet, according to one of the data owners: *"It would be useful to identify those data owners that have interesting data but are not yet aware of what open data is. They should be supported to open up their data."*

4.2 Preparation

After the three datasets to be opened up for the hackathon were identified, the second phase of the project consisted of preparing the datasets for publication. We found that although the datasets that were identified were of high quality, it still required some work before they could be opened up. Except for the involvement of the legal advisor, who checks whether the data that are to be made public can indeed be opened up, the main work in this phase was carried out by the information manager and the data owners. This phase consisted of two steps: setting the requirements, and (technically) preparing the data.

4.2.1 Setting requirements

In the first step of the preparation phase, the information manager and legal advisor formulated the requirements for the data to be opened. These requirements include technical requirements (such as data quality level, standards and metadata), economic requirements (such as value proposition and business model) and legal requirements (such as the license for re-use). Especially, the issue of data quality was addressed. The data owners were concerned that data quality would not be high enough. At the same time, they realized that the users (the participants in the hackathon) would probably not have the time to build a fully operational application based on the data. Therefore, it was decided that data quality would remain as it was. The data owners agreed that the desired data quality largely depends on the demands of users. For example, in the case of using real-time public transport data accuracy is of higher importance than when data of historic monuments are published. Still, a data owners wonders about who would be responsible for realizing re-use: *"I am not sure whether it is the task of the data owner to think about this. The central characteristic of open data is that the community is in a better position to think about re-use than data owners."*

4.2.2 Preparing the data

The second step of the preparation phase was the technical preparation of the data. This was the responsibility of the information manager and the data owner (or the person that is made responsible by the data owner). During this step a number of issues were addressed. Firstly, ownership of the data needed to be clear, otherwise data cannot be published freely. Secondly, data that can be tracked to individuals cannot be published or the part of the data that can be linked to individuals needs to be left out or anonymized. One of the data owners: *"We have had a discussion about privacy with the Dutch bureau of statistics (CBS), co-owner of the data. The Dutch data protection regulation is translated by CBS in rules on how to guarantee anonymity of the data, even when clever programmers work with the data."* Thirdly, data is often captured in an unstructured way that fits its original purpose. Therefore, this step included modelling the concepts and links within the data, and labelling the data in a unique way according to an Unique Resource Identifier strategy (similar to a website URL strategy). Fourthly, to allow re-use, data is converted into a machine readable and open structured format, metadata is added, and the data is stored following a specified format (e.g. SPARQL endpoints). Finally, we found that as there are no common protocols for opening up data yet, this provides the information manager with a lot of freedom, but also with many uncertainties.

4.3 Publication

The third phase of publication co-incided in this case study with its re-use: the data was published during a hackathon and instantly used by programmers to develop apps. We found that two steps were taken during the publication phase: ensuring technical findability and advertising the data. We found these two steps to have different purposes. While many organizations focus on the technical findability of data, also engagement with the community of potential re-users and advertising the data was found necessary to ensure data re-use.

4.3.1 Ensuring findability

During the first step of publishing the data, the data needs to be registered in such a way that they could be found by potential users. Technically, this was done by registering the data and metadata in the data catalogue of the hackathon, similar to publishing data in a national open data portal. The registration was found essential: it allowed data users to diminish the costs of finding the correct data. This task was carried out by the information manager and the data owners. While, in this case study publication of data took place in the confined environment of a hackathon, in a fully open environment this can be done in a national data portal. The data-owners stressed the importance of publishing documentation about the data: *"There was no description of our data uploaded on the servers, which made the data more difficult to re-use"*.

4.3.2 Advertising the data

While registration of the data in the most suitable portal and adding metadata may ensure findability, it may not be enough to actually ensure re-use. Ensuring proper advertisement of the data was the task of the community manager that communicated with potential users. During the hackathon the data owners presented the content of their data to potential re-users. Furthermore, the re-use conditions (license) were communicated to make sure that users understand them. Based on the reactions of the potential users we found this phase to be essential for realizing actual re-use. Merely publishing data is often not enough to show its potential. Still, it seems that many organizations forget this step and assume that opening up data as such is enough to ensure re-use. Data-owners need enough resources to advertise their data. One of the data-owners pointed out this problem: *"The hackathon is an initiative that is not part of my daily job. So, the hackathon was a personal hobby rather than a carefully planned project."*

4.4 Re-use

The fourth phase is the re-use of data. In the case study, however, we found that the data were not re-used during the hackathon – much to the dismay of the data owners. It seemed that the datasets that were opened did not respond to the wishes and interests of the teams of programmers. They stated that the data that the RTO opened up was often very complex and they could not easily grasp its potential during the one-day hackathon. Furthermore, there were many other datasets brought in during the hackathon. This meant that especially the step of advertising the data was essential to make sure that data would be re-used. What initially seemed to be a simple activity within the relative confined environment of a hackathon, thereby became a serious bottleneck in the process of opening up data. Having a community manager to guide the data owners through this step in the process is essential.

4.4.1 Building a community

This step was not actively undertaken during the process of opening up data of the RTO and its absence was felt as data were not re-used during the hackathon. The data owners found during the day that fostering re-use by building a community around the data that is opened up may be necessary to enable re-use. They said that besides advertising the availability of data during the data, the community manager could have actively sought to collaborate with external stakeholders that may want to use the data. This could have happened in earlier phases of the process. Active community building and involving external stakeholders already in the beginning of the process, may not only enable re-use, but may also spur the process of attracting feedback on the published data, which will help to improve the quality of the data. During the case study this appeared to be the major difficulty in ensuring data re-use. For example, the identification of the community and its stakeholders was very difficult, according to one of the data-owners: *"We tried to put ourselves in the shoes of the participants of the hackathon and found we could come up with a myriad of stakeholders."*

4.4.2 Managing the data

The responsibility of the information manager and the data owners was found not to stop after publication. Although the hackathon was finished after one day and the data were only published within the environment of this one-day activity, data owners still found that they needed to make a plan for how to manage the data and make sure that the data quality remains at the desired level. They also said that they would like to attract feedback from users, and that the information manager needs to be prepared for this, as well as for requests for support during re-use. At the moment this still takes place in an ad-hoc manner, according to one of the data owners: *"Open data does not have a place yet within our organization. It would be good if central servers would be installed especially for this purpose, that employees become responsible for it, and that the quality is managed."* In time, the RTO is considering to open up its own data portal instead of connecting with existing portals to allow for better management, feedback and support. Activities including regularly updates of the data and metadata, asking users for feedback to increase data quality, linking the data with new datasets within the community, and tracking visitors and users.

4.5 Evaluation

The last phase of the pilot project was the evaluation of the process of opening up data. While this was not a primary activity actually ensuring that data are opened up for the hackathon, it was found to be a crucial activity in the development of an open data strategy, spurred by the lack of re-use of the data that were opened up. Furthermore, during the fall of 2013 it was decided by the Ministry of Economic Affairs that the RTO needs to adopt an open data strategy for all research carried out using public funding. Hence, open data needs to become part of the organizational processes. To prepare for this process, an evaluation of the pilot project was considered necessary. All stakeholders were involved to see how open data can become embedded in the organizational strategy and work processes. Furthermore, the issue of community building to create more value from the datasets that are opened up was also addressed during the evaluation. The RTO considers open data not just as a 'compliance' issue that needs to be 'ticked off', but the organization feels the need to actively engage with the community that may want to use its data and support them in the process.

4.5.1 Assessing the data proposition

The first step of the evaluation phase consisted of assessing the value proposition of open data for the RTO. As described above, the RTO focuses strongly on enabling re-uses and aims to engage more actively with the communities that could benefit from re-use. In this phase, all relevant stakeholders were involved to determine the value of opening up data. Besides financial gains, also societal gains were considered. For example, the public interest in certain datasets is considered an important reason for opening up these data by the RTO. When the pilot project was finished, a new strategy setting step took place, when top management decided to roll out an open data strategy for the whole organization. It is expected that this strategy setting will again lead to a new cycle of the lifecycle. Thus, the process of opening up data likely requires multiple iterations. Another aspect of open data that is assessed is linking the data to other datasets, according to one of the data owners: *"Our data will become much more valuable if it is uploaded to a local portal such as the one owned by the city of Rotterdam, in order for users to experiment with combining datasets."*

4.5.2 Embedding the strategy in the organization and work processes

The last step of the evaluation phase is embedding open data in the organizational strategy and processes. This is mainly the responsibility of top management. In the case of the RTO this meant that the open data lifecycle was started anew, this time adjusted to fit the newer objective of opening up within the organization all data that can be opened that is funded by the government. This means that

on all organizational levels adjustments may be made to the work processes. As the pilot project already activated some data owners to actively become involved in communities around the data, we found that the project manager needs to balance innovations from top-down and bottom-up. A data owner: *"We need top-down as well as bottom-up support for open data. From a strategy to support our activities, to a well-supported data portal: it is all necessary."*

5 Findings and discussion

The findings from case study support the findings from literature and we found the refined open data lifecycle model also to consist of five phases. Based on the findings from the case study, the number of actions, however, was limited to two activities per phase, refining the activities of the lifecycle model of table 1 to a ten-step process. The main reason for this was to make sure that all steps of the lifecycle model have a similar level of detail. Furthermore, the case study identified a number of additional steps to be taken that were not found in literature. The differences between the existing and the refined open data lifecycle mode are shown in table 3.

Lifecycle phase	**Steps per phase (literature)**	**Steps per phase (case study)**	**Main differences**
Identification	Setting the strategy	Setting the strategy	--
	Selecting the data	Selecting the data	--
Preparation	Setting requirements	Setting requirements	--
	Modelling and describing data	Technically preparing the data	We found that existing models have a strong technical orientation, focusing on a lot of details this preparation phase.
	Converting to machine-readable data format		
	Linking data		
	Storing data		
Publication	Publication of data	Ensuring findability	The process of publishing entails more in practice than described by the models based on literature.
	Publication of metadata	Advertising the data	Rather than merely focusing on technical findability, actually advertising the data is necessary.
Re-use	Exploiting of published data	Community building	Exploitation requires active involvement with the community.
	Data management	Managing the data	--
Evaluation	Developing business propositions	Assessing the data proposition	The data proposition needs to be aligned with the strategic purpose of open data, and not just focus on compliance or re-use by others.
	Monitoring and improving data	Embedding into strategy and work processes	Balancing top-down strategy for open data and bottom-up initiatives (e.g. community building).

Table 3. Comparing the refined open data lifecycle model with existing models.

Similar to Ruparelia (2010) and Kalampokis et al. (2011), we found that the existing open data lifecycle models focus mainly on the technical aspects of opening up data without taking organizational aspects into account. Especially in the preparation phase, emphasis is put on all the

technical steps that need to be taken to ensure publication, but little attention is paid to involving relevant stakeholders. Especially to allow for re-use of data, involvement of external stakeholders such as potential users early in the process may enhance the chance of data re-use. The RTO in the case study currently focuses on community building with stakeholders outside the organization advertising the value of the data and ensuring re-use.

Furthermore, we found that the lifecycle models often focus on publishing data following the notion of compliance or making sure that they have opened up their data, and forget to follow the steps after the data has been published. Instead of focusing on supporting the process of re-use or evaluating the process, the existing lifecycle models merely envisage steps such as managing the published data, without considering the strategic importance of open data. In this way, opening up data remains an incidental process, rather than that it becomes an organizational routine to be applied to all relevant data within the organization. Thus, not only outside the organization, also within the organization relevant stakeholders, such as the community manager or the legal advisor are often not involved in the process of opening up data, while they are important for the development of an open data strategy.

Finally, we found that opening up data is iterative. Given the complexity, organizations likely go through multiple cycles to ensure optimal learning effects, or return to a previous step. This may help to gradually develop an open data strategy. Whereas the first cycle can serve as a pilot project for opening up a few datasets, a full strategy may be developed in a second cycle. A third cycle may be necessary in large organizations with laggards only included when the open data strategy is fully developed. Therefore, this model can serve as inspiration rather than as a prescription. Depending on where an organization currently finds itself in the process of opening up data, the lifecycle can be entered.

Further research should focus on how to align open data with the business proposition of organizations, as well as with the technology. While the lifecycle model that was developed in this paper is based on literature and practice, it was not validated for all organizations and open data strategies. Depending on the purpose of opening up data (e.g. compliance or commercial gains) the steps that are taken have different consequences. Therefore, to better understand the implications of the different open data strategies for the steps to be taken, the model should be validated using different types of organizations. Furthermore, this paper found that the process of opening up data should address technical as well as organizational aspects. Further research should thus look into how the interplay of the two takes place in practice and how they can be aligned during the process of opening up data.

6 Conclusion

Many government organizations are in the process of opening up their data to the public. However, they often find this process cumbersome and they do not know which steps to take. Lifecycle models for open data have been developed guiding the process of opening up data, but these are often not based on empirical findings. Therefore, we developed an open data lifecycle model based on literature and practice, using a case study of a semi-public organization in the Netherlands. Based on literature we identified five generic phases of opening up data: identification, preparation, publication, re-use, and evaluation. These were validated by the case study. Using the empirical findings, we then identified ten steps to be taken involving six different organizational stakeholders. We found that while most of the existing lifecycle models focus on the technical aspects of opening up data, the involvement of relevant stakeholders, both within and outside the organization is also essential to realize the support for the process and ensure re-use. Further research should look into the alignment of open data with other organizational goals as well as with the technology, to allow for a proper embedding of open data in the organization.

References

Alani, H., Dupplaw, D., Sheridan, J., O'Hara, K., Darlington, J., Shadbolt, N., & Tullo, C. (2007). Unlocking the potential of public sector information with Semantic Web technology. In: The 6th International Semantic Web Conference (ISWC), 11-15 Nov 2007, Busan, Korea.

Blanchard, B.S. & Fabrycky, W.J. (2006). Systems Engineering and Analysis, Fourth Edition. Prentice Hall. p. 19.

Curtin, G. G. (2010). Free the Data!: E-Governance for Megaregions. Public Works Management & Policy, 14(3), 307-326.

Ferrara, A., Genta, L., & Montanelli, S. (2012). Tailoring linked data exploration through inCloud filtering. In Proceedings of the 2012 Joint EDBT/ICDT Workshops (pp. 140-143). ACM.

Harrison, T. M., Pardo, T. A., & Cook, M. (2012). Creating Open Government Ecosystems: A Research and Development Agenda. Future Internet, 4(4), 900-928.

Hausenblas, M. (2011). Linked data lifecycles, presentation from DERI research institute, Galway, Ireland, July 2011.

Hyland, B. (2010). Preparing for a linked data enterprise. Linking Enterprise Data, 51-64.

Jaeger, P.T., & Bertot, J.C. (2010). Transparency and technological change: Ensuring equal and sustained public access to government information. Government Information Quarterly, 27(4), 371-376.

Janssen, M. & Zuiderwijk, A. (2012). Open data and transformational government, presented at the tGov conference, 8-9 May 2012, Brunel University, United Kingdom.

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

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2011). A classification scheme for open government data: towards linking decentralised data. International Journal of Web Engineering and Technology, 6(3), 266-285.

Klein, H. K. & Myers, M.D. (1999). A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems. MIS Quarterly 23 (1), 67-93.

Lane, S., & Richardson, I. (2011). Process models for service-based applications: A systematic literature review. Information and Software Technology, 53(5), 424-439.

McDermott, P. (2010). Building open government. Government Information Quarterly, 27, 401-413.

Mingers, J. (2001). Combining IS research methods: Towards a pluralist methodology. Information Systems Research, 12 (3), 240-259.

Ruparelia, N. B. (2010). Software development lifecycle models. ACM SIGSOFT Software Engineering Notes, 35(3), 8-13.

Stallinger, F., Neumann, R., Schossleitner, R., & Zeilinger, R. (2011). Linking Software Life Cycle Activities with Product Strategy and Economics: Extending ISO/IEC 12207 with Product Management Best Practices. Software Process Improvement and Capability Determination, 157-168.

Tsai, N., Choi, B., & Perry, M. (2009). Improving the process of E-Government initiative: An in-depth case study of web-based GIS implementation. Government Information Quarterly, 26(2), 368-376.

Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., & Gómez-Pérez, A. (2011). Methodological Guidelines for Publishing Government Linked Data. Linking Government Data, 27-49.

Walsham, G. (2006). Doing interpretive research. European Journal of Information Systems, 15(3), 320-330.

White Baker, E. (2010). Why situational method engineering is useful to information systems development. Information Systems Journal, 21(2), 155-174.

Yin, R.K. (1989). Case Study Research: Design and methods. Sage publications, Newbury Park, California.