

Walking the extra byte: A lifecycle model for linked open data

Auteurs

Tijs van den Broek (TNO)
Anne Fleur van Veenstra (TNO)
Erwin Folmer (TNO)

Abstract

More and more public organizations publish their data in an open format to increase transparency and foster economic activity. In this modern gold rush, organizations strive to open up as many datasets as possible, without considering the strategic importance of open data. Especially linking data to other datasets can lead to the creation of innovative services. One issue that is often not explicitly addressed before opening up data is the format of the dataset. Central to open data is that the format is machine readable. But to allow for effortless linking of datasets, data being merely machine readable is not sufficient. Lifecycle models can guide the process of publishing linked open data. Current linked open data lifecycle models focus on the technical steps that need to be taken by the internal IT organization and often forget to include actions to be taken after publication. The effectiveness of linked open data, however, depends on how much the data is used. Hence, this paper develops a linked open data lifecycle model that takes the multiple disciplines and stakeholders

within and outside the organization into account as well as the steps to be taken after publication of the datasets. Firstly, using existing linked open data lifecycle models, this paper identifies generic phases of opening up linked data: identification, preparation, publication, re-use and evaluation. Secondly, investigating the process of opening up data in a semi-public organization in the Netherlands, the lifecycle model is refined and detailed. This case study shows that the involvement of relevant stakeholders, both within and outside the organization and of various disciplines, is essential to realize the support for the process and stimulate re-use.

Keywords

Open government, Linked data, Open data, Lifecycle model, Case study

Linked open data as a strategic asset
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 driver for economic activity (Harrison & Pardo, 2012): the European Commission expects that the re-use of public sector information could an annual

economic impact of EUR 140 billion in the European Union (Vickery, 2011). Part of this economic impact comes from innovative services that combine two or more datasets. To allow for effortless linking of datasets, data being merely machine-readable is not sufficient. An extension of open data is linked open data. For linked open data, the semantics of the data are modelled and the data can be linked to and from external data sets (Bizer et al., 2009; Hausenblas, 2009), to allow government agencies to link their data while staying in control of their own data. The difference between open data and linked open data is clarified by the introduction of the 5-star classification of data (Berners-Lee, 2006). Figure 1 lists the five levels of data quality identified by Berners-Lee (2006).

Figure 1 The 5-star data model (Berners-Lee, 2006)



Available on the web (in whatever format) but with an open license, to be open data.



Available as machine-readable structured data (e.g. excel instead of image scan of a table).



As (2), plus: Non-proprietary format (e.g. CSV instead of excel).



All the above, plus: Use the open standards from the World Wide Web consortium (RDF and SPARQL) to identify objects in the data, so that people can refer to them.



All the above, plus: Link your data to other people's data to provide context.

Data classified with 1, 2 or 3 stars are typically termed open data, while data with 4 or 5 stars are termed linked open data. Linked open data is seen as a requirement for effective data re-use and hence an essential extension of open data. Linked open data has been championed by the British government since the publication of the Putting the Frontline First action plan in 2009. Despite the guidance by the internet scientists Tim Berners-Lee and professor Nigel Shadbolt, there are merely 200 linked open datasets available in the British data portal data.gov.uk (Huijboom & Van den Broek, 2011).

Organizations often find the process of opening up data burdensome (see

e.g. Janssen, Charalabidis & Zuiderwijk, 2012). They are often unaware which steps to take in the process of opening up linked data. Lifecycle models are used to guide this process. However, most of these linked open data lifecycle models focus on the technical steps that need to be taken by the internal IT organization and often forget to include actions to be taken after publication. Furthermore, most of these models focus 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. A strategic perspective of linked open data requires a process that involves internal and external stakeholders from a wide range of disciplines. Hence, this paper develops a linked open data lifecycle model that takes multiple disciplines and stakeholders within and outside the organization into account as well as the steps to be taken after publication of the datasets.

We develop a revised linked open data lifecycle model in two steps. Firstly, we assess current linked open data lifecycle models to identify generic phases that organizations opening up linked data go through. Secondly, 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. While the data in the case study was not published in a linked data format, the lessons learnt still add to current technically-oriented models. The next section presents existing linked open data 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 lessons from the case study by formulating the refined linked open data lifecycle model. Section five formulates conclusions and recommendations for organizations that are considering to open up their data.

An assessment of current linked open data lifecycle models
One way of structurally capturing challenges of information systems 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 describe the development of certain phenomena and predict the next steps in the development (Lane & Richardson, 2011). In contrast to maturity models, lifecycle models do not prescribe organizational stages of the software development process. We found seven lifecycle models describing the process of opening up linked data and guiding organizations through this process.

Table 1 gives an overview of these existing linked open data models and 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.

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

Lifecycle phase	Steps per phase	Activities in literature
Identification	Setting the strategy	Setting aims of linked open data (Alani et al., 2007)
		Data awareness (Hausenblas, 2011)
	Selecting the data	Deciding on making data available Janssen & Zuiderwijk, 2012)
		Collecting databases (Alani et al., 2007)
		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)

Lifecycle phase	Steps per phase	Activities in literature
Preparation	Setting requirements	Analysing requirements (Alani et al., 2007)
	Modelling and describing data	Specifying, defining and analysing the data (Villazon-Terrazas et al., 2011; Hyland, 2010; Hyland & Wood, 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; Villazon-Terrazas et al., 2011)	
Storing data	Storing data in a datastore (Ferrara et al., 2011)	
Publication	Publication of data	Publishing data (Hausenblas, 2011; Hyland, 2010; Hyland & Wood, 2011; Janssen & Zuiderwijk, 2012; Villazon-Terrazas et al., 2011)
	Publication of metadata	Publishing metadata (Villazon-Terrazas et al., 2011)
Re-use	Exploiting of published data	Creating an online data catalogue 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)	

Lifecycle phase	Steps per phase	Activities in literature
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)

Most of these models have been based on cases of linked open data in the public sector, focusing strongly on merely making sure that data are technically opened up to the public rather than ensuring that linked open data becomes part of the strategic mission of the organization. Therefore, we found that there is a need to develop a revised linked open data lifecycle using a case study of a semi-public organization aiming to embed linked open data in its strategy and work processes.

Opening up open data in a semi-public organization

Case study approach

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 (Netherlands Organisation for Applied Scientific Research), the national RTO of the Netherlands. This case was selected

as TNO 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 combine action research and semi-structured interviews. 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. The observations of the action research were validated 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. Interview 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. The findings from the desk research and case study result in a revised lifecycle model that formulate the steps and organizational stakeholders within each phase of the process.

Case description: TNO

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 up linked data was not undertaken in a structural manner, but took place incidentally. The RTO identified three different reasons to open up its data. 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 activi-

ties, 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 up to the public. During this pilot project three steps were taken. Firstly, suitable datasets that could be opened up were identified and the data owners of these datasets were invited to participate in this pilot. Three datasets were identified and subsequently prepared for opening up: traffic data, geological data and data on working conditions in the Netherlands. 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

A revised linked open data lifecycle model

To open up its data, the RTO took the steps visualized in the linked open data lifecycle model below (see figure 1). The model consists of five phases (identification, preparation, publication, re-use and evaluation), each consisting of two steps. Furthermore, the model distinguishes five organizational stakeholders: top management, information manager, legal advisor, community manager and data owner.

The model and the lessons learnt in the RTO case study are described step by step below.

Identification

The first phase of opening up data comprises the definition of the process of opening up data and the identification of data that are 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 consist of two steps: setting the strategy and identifying the data for opening up.

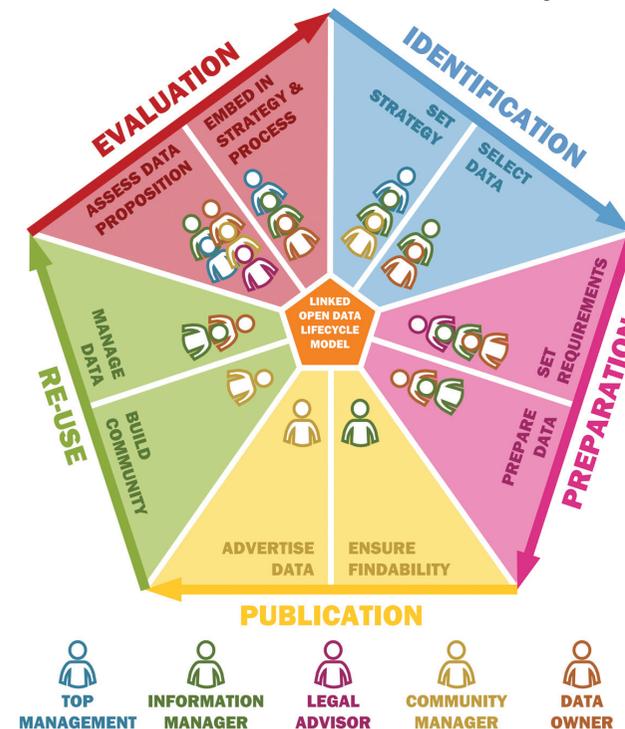


Figure 2 The revised linked open data lifecycle model

Setting the strategy

The first step in the identification phase is to develop a linked open data strategy. Top management should develop a vision on how linked open data contributes to the organizational mission.

A proper vision should not only include which data to publish, but also which data to re-use from others. Early top management support is of critical importance – even if linked open data merely starts off as a pilot project. While this may imply that a full strategy is not yet in place, it does mean that support is given to the process. In case the linked open data strategy includes fostering economic activity, in this phase also the connection with potential users may be useful to identify their requirements and demands.

Selecting the data

In the second step of the identification phase, the information manager and the data owners identify datasets that can be opened up, based on the linked open data strategy. Especially for larger organizations it is impossible to open up all available datasets at once. From a long list of available datasets that comply to the above-mentioned criteria, the most meaningful datasets should be selected: the shortlist. This selection can be based on developing a business case, in which the interest among users and the costs for opening up

the data is taken into account. This leads to a prioritized shortlist of datasets, on which to base the decision for the selection. This step also includes the mobilization of the data owners of the datasets on the shortlist.

Which datasets can be opened up?

- Datasets that are fully owned by the organization that publishes the data or for which a consent for publication has been obtained
- Datasets that do not contain classified information or information that contains data that is linked to national security
- Datasets that do not contain information that can be linked to individuals
- Datasets that are not exempted by third-party Intellectual Property rights or other exemptions formulated in the upcoming revision of the PSI directive

Source: EPSI platform, 2013

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 consists of two steps: setting the requirements, and (technically) preparing the data.

Setting the requirements

In the first step of the preparation phase, the information manager and legal advisor formulate the requirements of the data. 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 open license). Consequently, the project manager needs to involve all relevant all stakeholders in setting the data requirements to prevent any undesirable surprises later in the process. Depending on the linked open data strategy, the quality level requires more or less attention. Setting requi-

rements for the data quality includes assessing the current quality of data, setting goals for data quality and selecting data standards. Firstly, it is worthwhile to assess the quality of a dataset, the current level of 'stars'. Secondly, the desired level of stars for the data has to be set. When the goal is to publish linked open data, the desired level should be 4 or 5 stars. Based on the initial level of stars, each dataset needs a plan how to gradually reach the next levels towards linked open data. For example, when opening up new data without any stars, it is better to plan the first steps that aim for 1-3 star data, then directly go for 5 star data. Thirdly, the data standards need to be selected in advance. Linked open data helps to limit the selection of standards to the open semantic web standards (from W3C), such as RDF (RDF-S), and presented in open formats such as XML, N3, Turtle, and SPARQL to query the data.

It is also worthwhile to assess the intrinsic quality of a dataset with existing quality instruments (Folmer, 2012): how consistent, complete, reliable, etc. is the quality of the items in the dataset?

Preparing the data

How to prepare data for publication?

- Anonymizing any information that can be linked to individuals
- Modelling the concepts and links within the data
- Labelling the data in a unique way according to a Unique Resource Identifier strategy (similar to a website URL strategy)
- Converting data into a machine readable and open structured format (for 3-star data)
- Adding metadata
- Documenting the data for future re-use
- Storing the data following the four design rules for linked open data (for 5-star data)

The second step is the technical preparation of the data. This is the responsibility of the information manager and the data owner (or the person that is made responsible by the data owner) for managing a specific dataset. Depending on the data requirements set, this step includes modelling, description, conversion and storing of data. Firstly, ownership of the data needs 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. Thirdly, data is often captured in an unstructured way that fits its original purpose. Therefore, this step includes modelling the concepts and links within the data, and labelling the data in a unique way.

When preparing for linked open data the following design rules are advised (adapted from Berners-Lee, 2006):

- All elements in the dataset need to be uniquely identifiable by the use of Unique Resource Identifiers (URIs) as identifier, which is a strategy similar to URL. There are, however, many ways to construct a URI, and it is therefore preferred to adopt a naming convention. The concept URI strategy (for naming convention) for Dutch linked open (government) data is presented in this book (Brink, Overbeek & Bren-tjes, 2013), and is recommended to be used.
- Use HTTP URIs so that people can look up those names on the Internet.
- When someone looks up a URI, provide useful information, using open standards (e.g. RDF* and SPARQL) to provide this information.
- Include links to the URIs of other data sets so that data users can discover and link datasets.

- 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 (as defined in the requirements phase).

The converting and preparing of data sets is often combined with improving the intrinsic quality of the data, simply because during conversion many intrinsic quality issues will become apparent and needs to be solved.

Publication

The third phase of publication coincided 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.

Ensuring the findability

The first step of publication is to make sure that the published data can be found by users. This can be done by registering the data and metadata in an existing data catalogue, for example the national data

portal. Finding the right platform for publishing datasets is essential for attracting attention and users. This registration is essential: it allows data users to diminish the costs of data discovery. This job is done by the project manager and information manager. In a later stage (see step 8), you can consider to open up your own data portal, for example data.yourorganization.eu. Linked open data can improve the findability of the data: the URIs in the data are traceable, so the user can browse through the dataset, explore new related data sets and link to them. With having a starting point with some data, other related data can be discovered (Berners-Lee, 2006).

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. This is the task of the community manager, who can reach out using different forms of communication, such as press releases, blogs, app contests, hackathons, information days, or app awards. Furthermore, the re-use conditions (license) need to be communicated to make sure that users understand the conditions. The involvement of external stakeholders should be linked to the business case for selecting datasets in order to make sure that those datasets are opened that attract users.

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. Potentially, linked data solves this issue (third design rule as presented earlier: provide useful information about the data). 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.

Building a community

The first step in fostering re-use is building linked open data communities. Besides advertising the availability of data, the community manager should collaborate with external stakeholders in order to build an active network around your data. Stakeholders

can include civil rights organizations, web entrepreneurs, incubators, and research institutes. The community manager and legal advisor need to ensure that the technical, economic and legal requirements set in the preparation phase are implemented. The community manager should develop a plan that describes how to engage the right community, given the linked open data strategy from the beginning of the process. Active community building may also help the process of attracting feedback on the published data, which will help to improve the quality of the data.

Managing the data

How to manage linked open data?

- Regularly update the data and publish updates to ensure predictability
- Ask users to give feedback on data to increase data quality
- Update metadata
- Link data with new datasets within the community
- Track visitors and users

The responsibility of the information manager and the data owner does not stop after publication. They need to make a plan for how to manage the data and make sure that the data quality remains at the desired level.

The information manager needs to be prepared for receiving feedback from users, as well as requests for support during re-use. In time, organizations may even decide to open up their own data portal instead of connecting with existing portals to allow for better management and support.

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 (at least published under an open license) 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 considered 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.

Assessing the data proposition

The first step of the evaluation phase is assessing the value proposition of linked open data. In this step, the results of publication should be evaluated against the business case that was created earlier. Furthermore, the project manager should assess the impact of the published datasets using other indicators, such as the number of downloads, combinations with other datasets, users, applications and end-users of these applications. This assessment should be shared and evaluated with top management. The project manager may need to keep in mind that the value of open data is broader than merely financial benefits. For example, its social impact, such as increased transparency, can be more important than an increase in revenue – depending on the linked open data strategy that was formulated. It is expected that the evaluation may trigger strategy setting, which will again lead to a new cycle of the lifecycle. Thus, the process of opening up linked data likely requires multiple iterations.

Embedding the strategy in the organization and work processes

The last step of the evaluation phase is embedding linked open data in the organizational strategy and processes. Top management should follow up the lessons learned of the linked open data implementation in the organizational strategy, paying special attention to any changes in the organizational culture. This may mean an adjustment of the initial linked open data strategy. On the tactical level, the project manager should set practical guidelines for linked open data in the organizational processes. In this way, several steps of this linked open data process can be automated. The project manager and top management should balance innovation initiated top-down (implementing strategy) and bottom-up (encouraging new initiatives).

Conclusion

Many public organizations publish their data in an open format to increase transparency and foster economic activity. Most of these organizations strive to open up as many datasets as possible, without considering the strategic importance of open data: how does re-use add to the mission of the organization? To allow for effortless linking of datasets, data being merely machine readable is not sufficient. The standards for linked open data can foster the re-use of open data. The process of opening up linked data is seen as cumbersome and the number of linked open datasets is lacking behind. Lifecycle models can guide the process of publishing linked open data. Current linked open data lifecycle models focus on the technical steps that need to be taken by the internal IT organization and often forget to include actions to be taken after publication. The effectiveness of linked open data, however, depends on how much the data is re-used. Therefore, we developed a linked open data lifecycle model based on literature and practice, using a case study of a semi-public organization in the Netherlands. Firstly, we identified five generic phases of opening up linked data: identification, preparation, publication, re-use, and evaluation. These phases were validated in the case study. The case study shows that the involvement of relevant sta-

keholders, both within and outside the organization and of various disciplines, is essential to realize the support for the process and stimulate re-use. The resulting linked open data lifecycle model is developed based on the notion that a clear strategy needs to be in place to successfully open linked data. Currently, many organizations merely focus on compliance with open data regulation rather than they think about the strategic importance. A proper strategy determines choices such as which data to open up, which stakeholders to include, which data quality level to aim for, which portal to use for publishing the data, how to organize legal ownership, etc. While it can be very useful to learn from other organizations, it is even more important to determine what opening up linked data can do for the strategic goals of the organization. If innovation from data re-use is an important goal, it may pay off to identify potential users and their needs in the beginning of the lifecycle, and strive for linked open data to ensure effortless linking. Organizations, however, need to remain open to new opportunities, as the case study shows that it is hard to determine the full potential of data upfront.

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.
- Berners-Lee, T. (2006). Linked Data – Design Issues. Retrieved June 5, 2013, <http://www.w3.org/DesignIssues/LinkedData.html>
- Blanchard, B.S. & Fabrycky, W.J. (2006). Systems Engineering and Analysis, Fourth Edition. Prentice Hall. p. 19.
- Brink, van den, L. Overbeek, H., Brentjens, T. (2013). Designing A URI Strategy For The Dutch Public Sector, in Pilot Linked Open Data – Part 2. <http://www.pilod.nl/Boek/BrinkEtAl-URI>
- 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.
- Folmer, E. (2012). Quality of Semantic Standards, PhD Thesis, University of Twente.
- 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.
- Huijboom, N., Broek, T. van den (2011). Open data: an international comparison of strategies.
- European Journal of ePractice, Hyland, B. (2010). Preparing for a linked data enterprise. *Linking Enterprise Data*, 51-64.
- Hyland, B., & Wood, D. (2011). The Joy of Data-A Cookbook for Publishing Linked Government Data on the Web. In *Linking Government Data* (pp. 3-26). Springer New York.
- 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 eGov 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.
- 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.
- Stallinger, F., Neumann, R., Schosleitner, 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.
- Vickery, G. (2011). Review of recent studies on PSI re-use and related market developments. *Information Economics*, Paris.
- 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.

Auteur

Christophe Guéret (DANS, Free University Amsterdam)

There is an increasing interest towards publishing data open on the Web. Opening what used to be closed data brings a lots of opportunities in terms of social and economical development, participatory governance and improvement of research processes. Although the motivation is clear the publication is in practice open to a lot of different options. Even for one motivated in publishing his data online, it remains still unclear how to do it. This chapter discusses the wide range of data publication approaches and pays a particular attention to the Linked Open Data publication principles advocated by the W3C.

Publishing Open Data

According to the site opendefinition.org, Open Data is "A piece of data or content is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and/or share-alike.". This implies no specific way to publish the data, it also gives no clear definition as to what data is. Both points are left up to the discretion of the data publisher.

In order to share his open data on the Web, a publisher will first make his data ready and then put it somewhere where it can be found and downloaded.

Get the data ready

A first option for publishing data is the so called 'CSV' file. This simple and intuitive format defines a table with rows as data entries and columns as properties for the entries. The first row is typically used to indicate what the column contain and the first column is typically used as the identifier of the entry. Comas or tabs can be used to separate the cells, leading to Coma Separated Values document (CSV) in the former case and Tabular Separated Values document (TSV) in the later. Such documents can be opened by any spreadsheet software for easy manual manipulation and can also be parsed by most popular data processing software.

The main drawback of CSV documents is that their semantics is very poor - if not inexistent. The Simple Data Format SDF proposed by the OpenKnowledge Foundation is a way to tackle this by associating a JSON document to the raw CSV data. This additional document describes the type of the columns and provide some other meta data that make it easier to consume the CSV data. A similar approach can be found in the Data Set Publishing Language