OPENING UP DATA BEYOND COMPLIANCE

TNO innovation for life

More and more public organizations publish their data to increase transparency and foster economic activity. The process of opening up data, however, is often cumbersome. Many different stakeholders need to be involved in order to embed open data in the organizational strategy and processes. Based on practical experience, TNO developed an open data lifecycle model with five phases organizations need to go through to open up their data. This paper describes for each phase the subsequent steps to be taken and stakeholders to involve.

THE OPEN DATA CHALLENGE

Public organizations around the world are setting up open data strategies to increase transparency and foster economic activity. Open data increasingly becomes a 'must-have' for public organizations to become more accountable for their actions and strengthen citizen engagement. At the same time, open data is increasingly seen as a driver for innovation. Based on data gathered with public funds, businesses can develop new services using data published in an open and re-usable format. Currently, European regulation is being developed extending open data beyond government organizations to semi-public organizations, such as cultural heritage foundations,

public transport organizations and research institutes. The purpose of opening up data entails more than transparency or economic value: open data needs to contribute to their organizational strategy and operational activities.

Organizations often find the process of opening up data cumbersome. It involves many actions at the same time, including strategic alignment with the organizational mission, technical challenges to opening datasets and legal issues such as establishing ownership of the data. Furthermore, issues such as guaranteeing privacy and stimulating data re-use by others need to be resolved during

implementation of an open data strategy. Often, organizations aiming to open up their data focus on compliance to public pressure from regulation or public opinion, merely making sure that their data are published. Alternatively, they focus on short-term re-use by opening up their data to hacker communities or by sponsoring app contests. While both of these pathways lead to open data, neither of these realize the envisaged impact: transparency or innovation. We argue that successful implementation of an open data strategy requires a more structured approach. To guide this approach we present an open data lifecycle model.

A LIFECYCLE MODEL

To successfully open up their data, (semi-) public organizations need to take a number of steps: the open data lifecycle model (see figure 1). This model was developed using a combination of desk research and action research. While opening up our own datasets, we evaluated each phase of the process, formulating which steps to take and which stakeholders to involve.

The open data lifecycle model consists of five phases (identification, preparation, publication, re-use and evaluation), each consisting of two steps. Besides the project manager, who is involved in every step, the model distinguishes five additional organizational stakeholders: top management, information manager, legal advisor, community manager and data owner (see checklists Which stakeholders to involve?). The model is described step by step.

Which stakeholders to involve?

✓ *Project manager: to manage the whole open data process*

✓ *Top management: to ensure that the necessary support is given to the process and to embed open data into the organizational strategy*

✓ *Information manager: to ensure that the data are prepared before and managed after opening up*

✓ *Legal advisor: to help determine data ownership and make sure that only those datasets are opened up that can be made public*

✓ *Community manager: to actively involve internal and external stakeholders, and introduce their needs into the process*

✓ *Data owner: to make sure that the dataset is kept up-to-date, also after opening up*

IDENTIFICATION

The first phase is the identification of data that will be published. In order to identify meaningful datasets, first the strategic proposition of opening data is formulated. Secondly, suitable datasets are identified (see checklist Which datasets can be opened up?).

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*
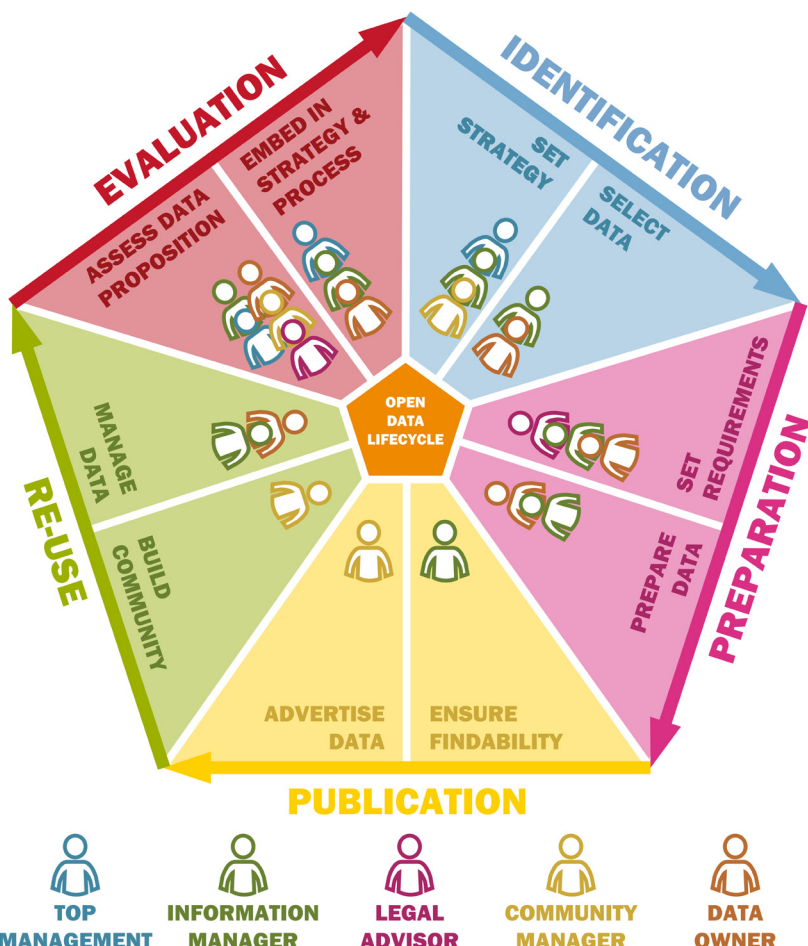
Source: EPSI platform, 2013

1. SET STRATEGY

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

2. SELECT 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 open data strategy. Especially for larger organizations it is impossible to open up all available datasets at once. From a longlist 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

Figure 1. The open data lifecycle model

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.

PREPARATION

During the second phase, the datasets are prepared to be published. Data is rarely immediately ready to be published and many issues need to be addressed in this phase (see the textbox How to prepare data for publication?).

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 an Unique Resource Identifier strategy (similar to a website URL strategy)*

- ✓ *Converting data into a machine readable and open structured format*

- ✓ *Adding metadata*

- ✓ *Documenting the data for future re-use*

- ✓ *Storing the data following a specified format, e.g. SPARQL endpoints*

3. SET 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). Furthermore, the issue of data quality needs to be addressed. Depending on the open data strategy, the quality level is more or less important. 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.

4. PREPARE 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. Furthermore, this includes ensuring that the data have the desired data quality. 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. Or in case the purpose of opening data is to spur participation, mistakes may be corrected and information can be added by citizens.

PUBLICATION

The third phase is the actual publication of the data.

5. ENSURE 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.

6. ADVERTISE 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 re-use of data. This phase includes leveraging re-use and managing the data (see How to manage open data?).

How to manage 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*

7. BUILD COMMUNITY

The next step is fostering re-use by building 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 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.

8. MANAGE DATA

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 is the evaluation of the open data process.

9. ASSESS DATA PROPOSITION

The first step of the evaluation phase is assessing the value proposition of 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, 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 open data strategy that was formulated.

10. EMBED IN STRATEGY & PROCESS

The last step of the evaluation phase is embedding open data in the organizational strategy and processes. Top management should follow up the lessons learned of the 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 open data strategy. On the tactical level, the project manager should set practical guidelines for open data in the organizational processes. In this way, several steps of this 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).

MORE THAN ONE CYCLE

The next phase of the lifecycle model could be to go back to the beginning; 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 (see figure 2).

Figure 2. Opening up data can require multiple iterations

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 being 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.

IMPLICATIONS FOR YOUR ORGANIZATION

The open data lifecycle model is developed based on the notion that a clear strategy needs to be in place to successfully open 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, which stakeholders to include, which portal to use for publishing the data, how to organize legal ownership, etc. While it can be very useful to learn from others, it is even more important to determine what opening up data can do for your organization. If innovation from data re-use is an important goal, it may pay off to identify the potential users and their needs in the beginning of the lifecycle.

Remain open to new opportunities, though! It is hard to determine the full potential of your data upfront.

TNO.NL