EUROPEAN
DATA PORTAL

Open Data Goldbook for Data Managers and Data Holders

Practical guidebook for organisations wanting to publish Open Data

Open Data Goldbook for Data Managers and Data Holders
Practical guidebook for organisations wanting to publish Open Data

Last update: January 2018

www   http://www.europeandataportal.eu
@   info@europeandataportal.eu
Licence: CC-BY

Capgemini Consulting prepared this Goldbook as part of the European Data Portal project. The European Data Portal is developed by the European Commission with the support of a consortium led by Capgemini Consulting, including INTRASOFT International, Fraunhofer Fokus, con.terra, Sogeti, the Open Data Institute, Time.Lex and the University of Southampton.

For more information about this Goldbook, please contact:

European Commission
Directorate General for Communications Networks, Content and Technology
Unit G.1: Data Policy & Innovation
Daniele Rizzi – Policy Officer
Email: Daniele.Rizzi@ec.europa.eu

Project team
Dinand Tinholt – Vice President, EU Lead, Capgemini Consulting
Executive lead European Data Portal
Email: Dinand.Tinholt@capgemini.com

Wendy Carrara – Principal Consultant, Capgemini Consulting
Project Manager European Data Portal
Email: Wendy.Carrara@capgemini.com

Written and reviewed by Wendy Carrara, Sem Enzerink, Fréderique Oudkerk, Cosmina Radu, Eva van Steenbergen (Capgemini Consulting)

DISCLAIMER

Table of Contents

EUROPEAN
DATA PORTAL

Reading guide

Different studies address the benefits and expected impact of sharing data. We have all seen the colourful infographics picturing the number of terabytes of data created over the internet each second. How does government data fit into the picture? The publication of the G8 Open Data Charter, among other global initiatives, has further underlined the value of opening government data. The re-use of Public Sector Information has been on Europe's agenda for over a decade now. Beyond adopting legislation, accessibility to data must continue and accelerate its pace. Nonetheless, publishing Open Data is a challenging task. It affects the technical data infrastructure, as well as organisational processes throughout the Open Data value chain: from data collection to the creation of new services and products.

As a first step, it is important to understand Open Data, to underline its benefits and to learn the (technical) concepts that accompany data publishing. Open Data is technically and legally open data that is accessible to everyone, can be manipulated, re-used and redistributed by anyone, for any purpose.

To support organisations on their path to 'open by default', the Open Data Goldbook for Data Managers was created. This Goldbook is a summary of all that you need to know as a Data Manager or Data Holder to implement an Open Data initiative successfully within your organisation. Public sector organisations, as well as the private sector can benefit from material compiled within this guidebook. From terminology to processes and from implementation to execution; you can expect this guidebook to cover the basic organisational, technical, and day-to-day challenges. Links to our eLearning modules are provided for you to dive deeper into the content, where you feel it may be necessary. Question you may have at this point:

| Why publish Open Data? | What does it take to publish Open Data? | What about regulation? | What technical solutions? | Where do I start? |
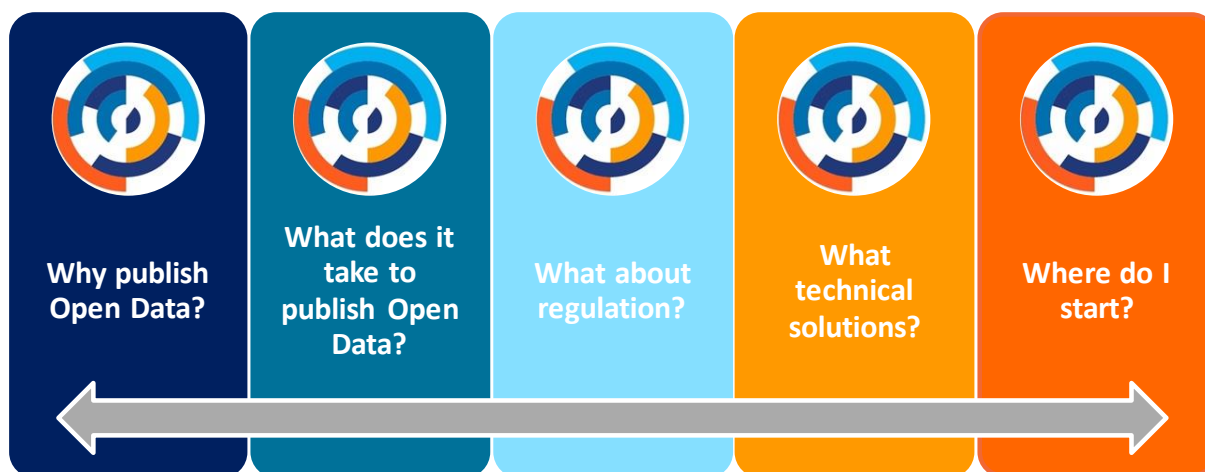
You will encounter different terms related to "Data" within this book. You might come across Open Data, Linked Data, Open Government Data, Public Sector Information, and even a mix of those. Be aware that, even if they seem to be equal, there are slight differences between them. Please refer to the Glossary for a complete overview of all terms applied.

Even though you might encounter several challenges while starting your Open Data initiative, such as internal resistance, challenges with drafting a policy and technical restrictions, at the end of this book

you will feel equipped with the basic knowledge that will help you face the crossroads and making the necessary choices.

Within the Goldbook, you will read about: Open Data in a Nutshell, How to build an Open Data Strategy, Technical preparation and implementation, Putting in place an Open Data lifecycle and Ensuring and monitoring success. The Goldbook consists of five chapters:

1 **Open Data in a nutshell**	1.1 General definition of Open Data 1.2 Public Sector Information & Open Government Data 1.3 The benefits of Open Data
2 **How to build an Open Data strategy**	2.1 Setting the ambition 2.2 Create your strategy 2.3 Drafting an Open Data Policy
3 **Technical preparation and implementation**	3.1 Data management 3.2 Extract, transform, publish 3.3 Channels 3.4 Search 3.5 Pre-requisites, choices and accountability
4 **Putting in place an Open Data lifecycle**	4.1 Collecting the data 4.2 Preparation 4.3 The Final Check
5 **Ensuring and monitoring success**	5.1 Engaging re-users 5.2 Monitoring your Open Data initiative

The Open Data Goldbook for Data Managers can be downloaded from the European Data Portal website. It is published under an Open Licence. Feel free to redistribute this Goldbook.

Various actors, so-called personas, have different roles to play when it comes to designing and implementing an Open Data initiative. In addition, not everyone knows where to start nor has a clear picture of what aspects need to be addressed. Different roles come into play. One might have to write a policy, the other might have to develop a portal, and another may collect data. To address the different roles involved in Open Data, the Open Data Goldbook was developed, introducing 4 roles ("Personas") in the Open Data Journey. This document briefly introduces these personas and their journey.

EUROPEAN
DATA PORTAL

Typically, there are four different personas involved in publishing Open Data:
- The Decision Maker
- The Data Manager
- The Developer
- The Contributor

These four key personas are introduced briefly below.

Decision Maker

The Decision Maker typically is a political figure who is responsible for a department, city, or maybe even a country. He or she is not particularly responsible for data, but can be the main sponsor of the Open Data strategy. He or she will validate the overall approach, oversee the implementation of the Open Data initiative and is ultimately accountable for the Open Data strategy.

The Decision Maker is not particularly involved with the technical topics regarding Open Data, as long as the data is published, and the IT requirements are managed within budget. His of her typical interest lies in understanding the benefits of implementing Open Data and getting started with an Open Data initiative.

Data Manager or Data Holder

A Data Manager or Data Holder is someone within an organisation who is responsible for collecting and sharing the data, regardless of whether an Open Data policy has been set up. This can be a manager, or someone who is dedicated to the topic. When launching an Open Data initiative, the Data Manager will be responsible for designing and implementing the Open Data strategy. This person has to know everything: from benefits, to barriers and from organisational choices to detailed publication instructions. He or she should be equipped with all knowledge available to implement the Open Data strategy successfully.

The Data Manager is interested in both organisational as well as technical topics. He or she might encounter resistance within his organisation and therefore should know how to overcome this. Furthermore, he or she should know about legislation, technical aspects, day-to-day instructions for employees who work with data, and more.

Developer

The Developer is typically responsible for implementing the technical requirements. Knowledge about technical standards, specific tools, as well as basic organisational requirements is therefore necessary. The Developer can be either an internal or an external resource assigned by the Data Manager. The two actors will actively collaborate.

Contributor

The Contributor can be any civil servant or contractor who works with data within a given (public) organisation. When the Open Data strategy is implemented, the Contributor will have an active role in collecting, preparing, publishing, and maintaining the data. The Contributor should be aware of the policies of the organisation and needs to know the standards.

EUROPEAN
DATA PORTAL

This Goldbook contains specific highlights in an easily readable fashion. In this Goldbook, you will find:

Quotes:

"Example quote for the Technical preparation and implementation section"

Recommendations:

> **This is a typical recommendation from the Technical preparation and implementation section**

Best Practices:

> A typical Best Practice from the Open Data in a Nutshell section

Case Studies:

> A typical Case Study from Technical preparation and implementation section

We also encourage you to consult the information provided in **"Appendix 6 - Online training material "** and the 16 online training modules included on the European Data Portal:

eLearning

Welcome to the European Data Portal's eLearning programme. Our online content gives you a simple, clear introduction to Open Data.

What is eLearning?

Our experts have selected 13 short modules designed for anyone to discover more about Open Data. The modules suit all levels from beginners to experts.

How long will this course take?

We realise that most people are too busy to take a whole course at once so our modules have been conceived to fit even into a busy schedule. Each module can take between 15 and 30 minutes to complete for those who wish to get just an overview. Each module also provides up to 2 hours of extra reading for those who want to dive deeper into a specific area.

When to use the modules?

For best results, complete all 13 modules following their order for a solid grounding in all aspects of Open Data. Or, if you prefer to pick a specific topic, feel free to proceed straight to this module.

What can you learn?

The eLearning programme introduces you to every aspect of Open Data. You will learn about definitions related to the concept and read success stories from across Europe. We introduce the major trends in Open Data, explain how people publish, access, and use it. Finally, we highlight the future of Open Data and get you thinking about the next steps for your own work.

Where do I go?

To get started click the link.

http://europeandataportal.eu/elearning/en/#/id/co-01

EUROPEAN DATA PORTAL

Glossary

API Application Programming Interface. A software intermediary that allows for distinct applications or systems to interact with one another

Bulk Download A download that contains multiple ranges (e.g. multiple time frames) of data and can be selected and retrieved at once

Buy-in An agreement on a policy or suggestion

CoE Centre of Expertise

CKAN Comprehensive Knowledge Archive Network. Open source catalogue system

Data Portal A software solution (usually a web site) that presents a catalogue of searchable and downloadable datasets in a user-friendly and uniform way. In general, each information source gets a dedicated web page

DCAT (-Application Profile) Data Catalogue Vocabulary, is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web. This document defines the schema and provides examples for its use. The Application Profile (-AP) is developed by the EC for interoperability optimisation between European Data Portals

EC European Commission

ETP-Process Extract, Transform, Publish- process. The process that starts with (raw) data in a database and ends with a publishable, published dataset

EU European Union

G8 Group 8: the leaders of 8 advanced economies in the world: The USA, The UK, Canada, Italy, Germany, France, China and Russia

Harvesting Web scraping. Computer software technique of extracting information from websites

Linked Data A method of publishing structured data so that it can be interlinked and become more useful through semantic queries, facilitating the sharing of machine-readable data on the web to be used by public administrations, business and citizens

Machine-readable A form of data that a computer can process

Metadata Data about data

OGD Open Government Data. Public Sector Data that has been published as Open Data

Open Data Data carrying an open licence stating it can be freely used, re-used and redistributed by anyone, for any purpose

Open Data Lifecycle The process of collecting, preparing, publishing, and maintaining Open Data

Policy A course or principle of action adopted or proposed by an organisation or individual

Proprietary Format A file format that is bound to proprietary software

PSI Public Sector Information: any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording) when produced by a public-sector body within its mandate

RDF Resource Description Framework: a standard model for data interchange on the web

RDFa An extension for embedding RDF

Re-user A person or organisation that uses existing (Open) Data for their purposes

Licence A legal permit to do something. A data owner should provide a licence with the data to specify the allowed re-use of the data

Interoperability The ability of different information technology systems and software applications to communicate, exchange data, and use the information that has been exchanged. For Data Portals this means a uniform way of publishing data

URI Unique Resource Identifier: string of characters used to identify a resource.

W3C The World Wide Web Consortium: global community responsible for developing web related standards

Web Publication Data published on a website

EUROPEAN
DATA PORTAL

Open Data Goldbook for Data Managers and Data Holders

Practical guidebook for organisations wanting to publish Open Data

EUROPEAN
DATA PORTAL

1. Open Data in a Nutshell

What is Open Data exactly? Various explanations exist. This section will offer a series of definitions. Furthermore, we will explain the differences between Open Data and PSI as well as Open Data and Open Government Data. Finally, we briefly explain why Open Data matters and what benefits can be expected.

Figure 1: Topics discussed in this chapter

1.1. General definition of Open Data

To provide a single definition of Open Data, we refer to the Open Definition (Open Knowledge, 2015) published by Open Knowledge. Open is in this case:

> *"Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness."*

The term Open Data is very specific and covers two different aspects of openness:

- 🌐 The data is legally open, which in practice generally means that the data is published under an open licence and that the conditions for re-use are limited to attribution.
- 🌐 The data is technically open, which means that the file is machine readable and non-proprietary where possible. In practice, this means that the data is free to access for everybody, and the file format and its content are not restricted to a given non-open source software tool.

Open Data can be freely used, modified, and shared by anyone. These properties are described in a licence. The fact that Open Data can be freely used in different ways does not necessarily mean that the data is available without charge. However, it is preferably downloadable via the Internet without charge (Open Knowledge, 2015).

To learn more about what Open Data is, please also refer to the relevant Online Training module about this: http://europeandataportal.eu/elearning/en/module1/#/id/co-01

Do you know what Open Data is yourself, but want to provide training to your colleagues? Go to the Training Companion that helps you deliver training:

https://www.europeandataportal.eu/en/resources/training-companion

1.2. Public Sector Information and Open Government Data

Public Sector Information (PSI) is the wide range of information that public-sector bodies collect, produce, reproduce, and disseminate in many areas of activity while accomplishing their institutional tasks. PSI may include (among others) social, economic, geographical, cadastral, weather, tourist, and business information.

Open Government Data refers to the information collected, produced or paid for by the public bodies (PSI) and made freely available for re-use for any purpose. Open Government Data is published under an open licence and is free to use within private and public domains.

In 2013, the G8 summit defined the importance of Open Government Data by creating the Open Data Charter. This charter emphasises the role that Open Data can play in both governance and growth stimulation. The charter defines five principles that nations that open their data should follow. The 5 basic principles are shown in Figure 2.
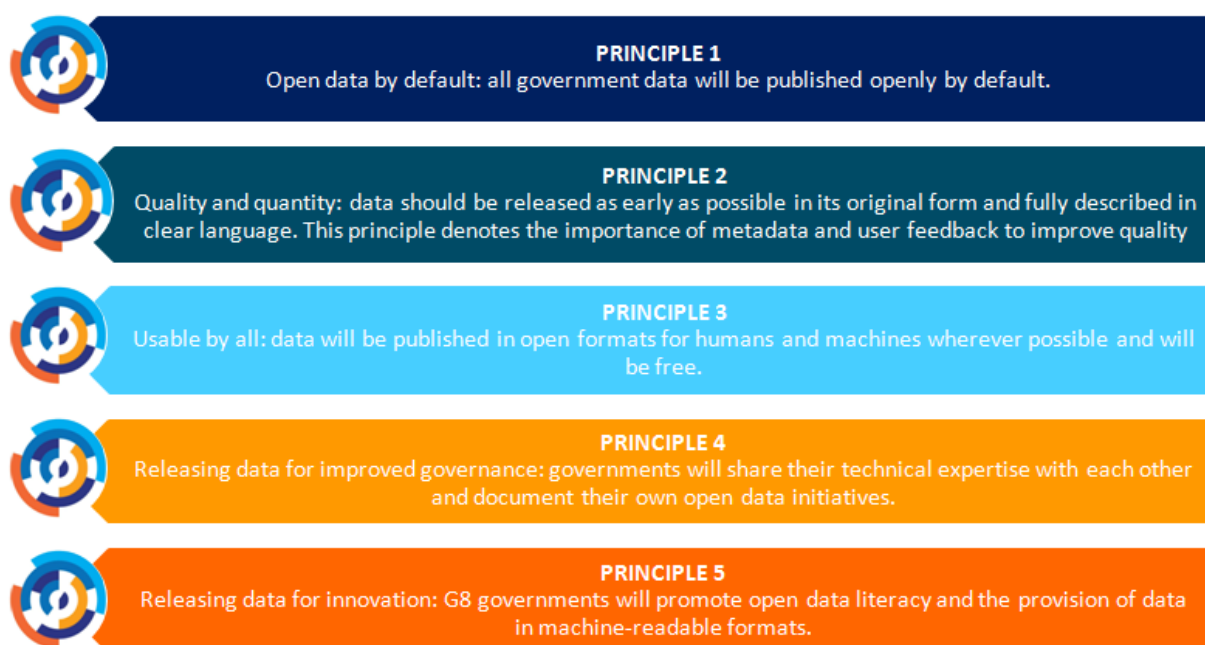

PRINCIPLE 1
Open data by default: all government data will be published openly by default.

PRINCIPLE 2
Quality and quantity: data should be released as early as possible in its original form and fully described in clear language. This principle denotes the importance of metadata and user feedback to improve quality

PRINCIPLE 3
Usable by all: data will be published in open formats for humans and machines wherever possible and will be free.

PRINCIPLE 4
Releasing data for improved governance: governments will share their technical expertise with each other and document their own open data initiatives.

PRINCIPLE 5
Releasing data for innovation: G8 governments will promote open data literacy and the provision of data in machine-readable formats.

Figure 2: The Open Data Charter Principles

The Directive on the re-use of Public Sector Information (2003/98/EC), also known as the PSI Directive, provides a common legal framework for a European market for government-held data. Therefore, within the European Union – thanks to this PSI Directive – PSI acquires a specific legal meaning with a framework providing a minimum set of requirements. A revision of the PSI Directive was introduced in 2013 (2013/37/EU). The main amendments are the breakaway from cost-based charging for PSI towards a margin-oriented fee, inclusion of certain cultural institutions as public-sector bodies, increased transparency regarding calculation of the fees, and support to machine-readable and open formats. The European Commission is currently reviewing the PSI directive, fulfilling its periodic review obligation, as part of a more general set of initiatives further implementing the Digital Single Market Strategy in the field of the European data economy.

1.2.1. Is PSI Different from Open (Government) Data?

While the terms PSI and Open (Government) Data are used quite often without distinction (thus overlapping most of the times), a strict definition of PSI according to the PSI Directive would reveal certain discrepancies. One should keep in mind that both the PSI directive and the so-called Open Data Movement provide a set of rules and principles that may be practically implemented in a slightly different way within different countries and different existing legal frameworks. There are several distinguishing arguments that are further explained in **"Appendix 1 - The PSI Directive vs. Generally Acknowledged Open Data Features"**. In summary, the main difference is that PSI refers to data held by public sector bodies only, and that its re-use may, under certain circumstances, be charged for. If PSI is made available under an open licence, it is called Open Government Data. The general term Open Data also refers to other types of non-public-sector data that is freely available, for example social media data.

Figure 3: From PSI to Open Data

1.3. The benefits of Open Data

Open Government Data is a wealth of untapped potential. As with any initiative within the public domain, it also involves expenditures and the effort of internal resources. Better understanding the benefits of Open Data can help accelerate the commitment around your Open Data initiative. The following overview provides more evidence of these benefits to support your initiative.

PSI that is generated and collected on a regular basis has a tremendous potential. The economic potential of PSI within the European Union goes beyond billions of Euros annually, with a potential to stimulate the overall economy and create new jobs. The economic study done by the European Data Portal estimates the direct market size of Open Data to be 325 bn. EUR for the period 2016-2020 and the number of Open Data jobs up to 100,000 in 2020. For more facts and figures on the economic benefits of Open Data visit the 2017 EDP report on The Economic Benefits of Open Data, available at: https://www.europeandataportal.eu/sites/default/files/analytical_report_n9_economic_benefits_of _open_data.pdf

EUROPEAN
DATA PORTAL

Making information that is generated and collected by public sector entities available and re-usable is important for many reasons:

- It provides citizens with a reliable knowledge base regarding government and public sector bodies' activities.
- It enables them to take part in public sector bodies' activities and therefore participate actively to the public choices (eDemocracy).
- It represents the initial material for public or private stakeholders to develop new added-value services and supply them to citizens.
- It is one of the crucial tasks to fulfil the aim of the Digital Agenda for Europe to "deliver sustainable economic and social benefits from a digital single market based on fast and ultra-fast internet and interoperable applications" (Kolodziejski, 2013)

Practically speaking, the benefits of Open Government Data may differ according to the type of stakeholder involved. These stakeholders can be divided into 3 main groups: governmental organisations, citizens and re-users. We will elaborate on various benefits for each of these stakeholder groups.

Firstly, governments themselves are one of the main re-users of the data they collect themselves. Practice has proven that by publishing data, governments themselves start re-using it, which results in costs savings (Berners-Lee, 2015). The following quote illustrates this example:

"This is why it should come a surprise that when the government of British Columbia began releasing Open Data in a centralised place, their Open Data portal, around one third of the site visits came from within government itself." - Rogers, 2015

Publishing Open Data enables the sharing of information within governments in machine-readable interoperable formats, which results in reducing costs of information exchange and data integration, no or limited upstream data management, error reduction by having one copy instead of multiple ones, etc. This results in improved data management, in terms of both quality and efficiency, as well as an overall reduction in administrative costs. In fact, the Greater Manchester area has estimated that freedom of information requests cost public bodies over £4 million a year, while over 600 public officials a day are unable to find or use data that they require for their jobs, costing authorities over £8.5 million a year. By breaking down the silos that exist between the various departments, bodies, and layers of government and allowing a fluid data flow can have substantial efficiency gains.

The economic analysis conducted by the European Data Portal estimates the accumulated cost savings for the EU28+ in 2020 to equal 1.7 bn. EUR

There are further benefits to consider:

- **Opening up data can optimise your process internally.** When data is open, none of your colleagues will have to go through an internal process to receive particular data. Many organisations have encountered the benefit of having their data open, simply because it takes less time to find data. Remember, your organisation will most probably be the most active re-user of your data.
- **Not only your organisation, but also citizens will benefit from an improved – and perhaps faster – internal information structure**. Processes will take less time, services can be digitalized, and citizens will benefit from more efficiency and transparency. A simple example might be to apply a single data provision to your services, thereby ensuring that

users – citizens and / or businesses – will not have to keep on providing data you already have.

- **If your organisation's data infrastructure may be outdated, your Open Data initiative might be a wonderful chance to achieve an internal change**. Many organisations have taken the opportunity to redesign their internal data infrastructure and incorporated the publication of data as a main activity in working instructions. Talk with the managers within your organisation what the plans are concerning IT infrastructure on data level.
- **By means of user feedback, you can improve the quality of your datasets.** The power of the crowd, known as crowd sourcing, is a very efficient way of pooling resources to reach a given, sometimes surprising, result.

> Transport of London benefits from a longer tradition of publishing data. They estimated that their data gains £130m annually. Read the full story: https://www.europeandataportal.eu/en/news/gaining-%C2%A3130m-annually-realising-london-transport-data

Secondly, by publishing PSI, government actions become more visible. This type of transparency helps both governments and citizens. It enables citizens to verify government actions. In turn, citizens' understanding of government will increase and they will feel more empowered by increased access to information. This empowerment could stimulate democracy and participation in (local) government.

Some examples of the effects of these transparency benefits are spending transparency, such as agricultural subsidies and the press coverage of those data, and elections results (e.g. EP elections or national/regional elections) in EU Member States. For citizens there are several benefits, such as the afore-mentioned transparency, but also possible social and commercial value. Furthermore, as citizens are better informed, they can actively participate and cooperate with the (local) government.

Besides the creation of social value, Open Data opens up possibilities for entrepreneurs. Open Data creates value for both citizens and private businesses after the release of a specific application. Social value for the public sector can generate commercial value for the private sector. Data is a key resource and as such, Open Government Data has tremendous commercial value. Since governments typically hold large amounts of information stored in all kinds of systems, opening up this data would lead to freeing up this potential. Hand in hand with Big Data, Open Government Data stimulates re-users to create new innovative products and services. Innovation is a key driver of long-term commercial success and stimulated by Open Data. It has a large potential to stimulate economic growth. In addition, the re-use can stimulate the improvement of processes, such as planning, quality and digitalisation. For some businesses, this means an in-depth transformation of business models and therefore internal innovation can equally be achieved.

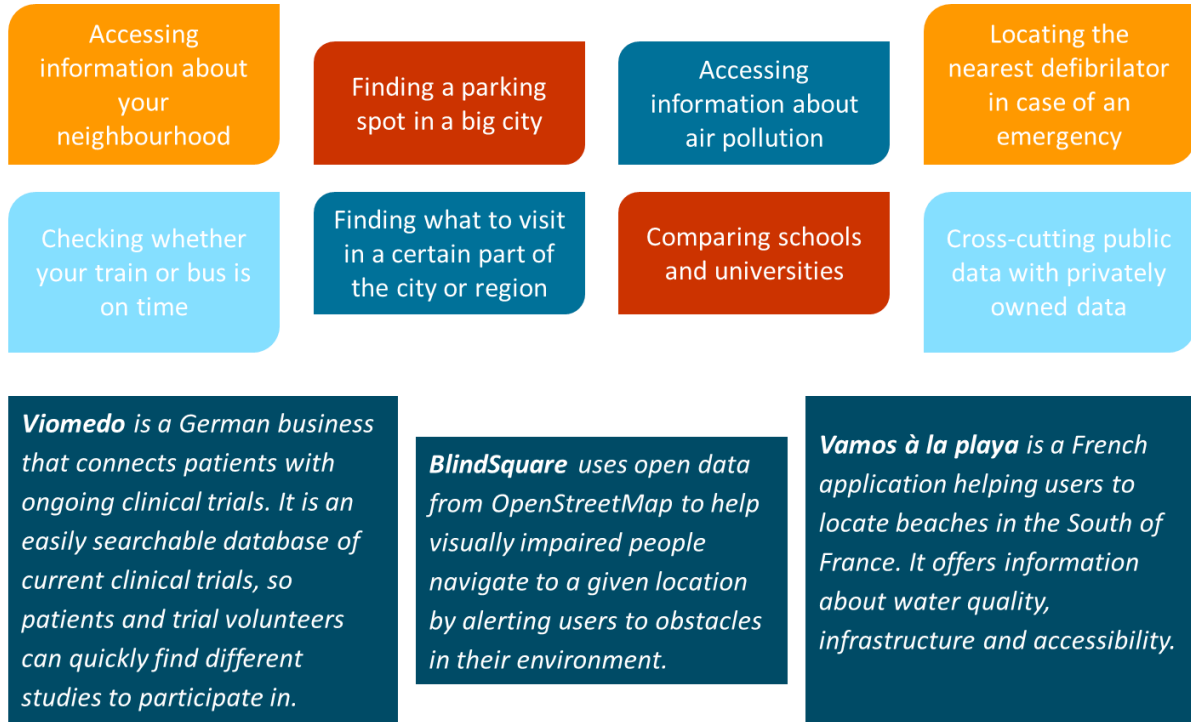There are numerous examples of re-use of Open Data as shown below:

Accessing information about your neighbourhood	Finding a parking spot in a big city	Accessing information about air pollution	Locating the nearest defibrilator in case of an emergency
Checking whether your train or bus is on time	Finding what to visit in a certain part of the city or region	Comparing schools and universities	Cross-cutting public data with privately owned data

Viomedo is a German business that connects patients with ongoing clinical trials. It is an easily searchable database of current clinical trials, so patients and trial volunteers can quickly find different studies to participate in.

BlindSquare uses open data from OpenStreetMap to help visually impaired people navigate to a given location by alerting users to obstacles in their environment.

Vamos à la playa is a French application helping users to locate beaches in the South of France. It offers information about water quality, infrastructure and accessibility.

Figure 4: Examples of re-use of Open Data

The following figure summarises the main benefits presented in the previous paragraphs:
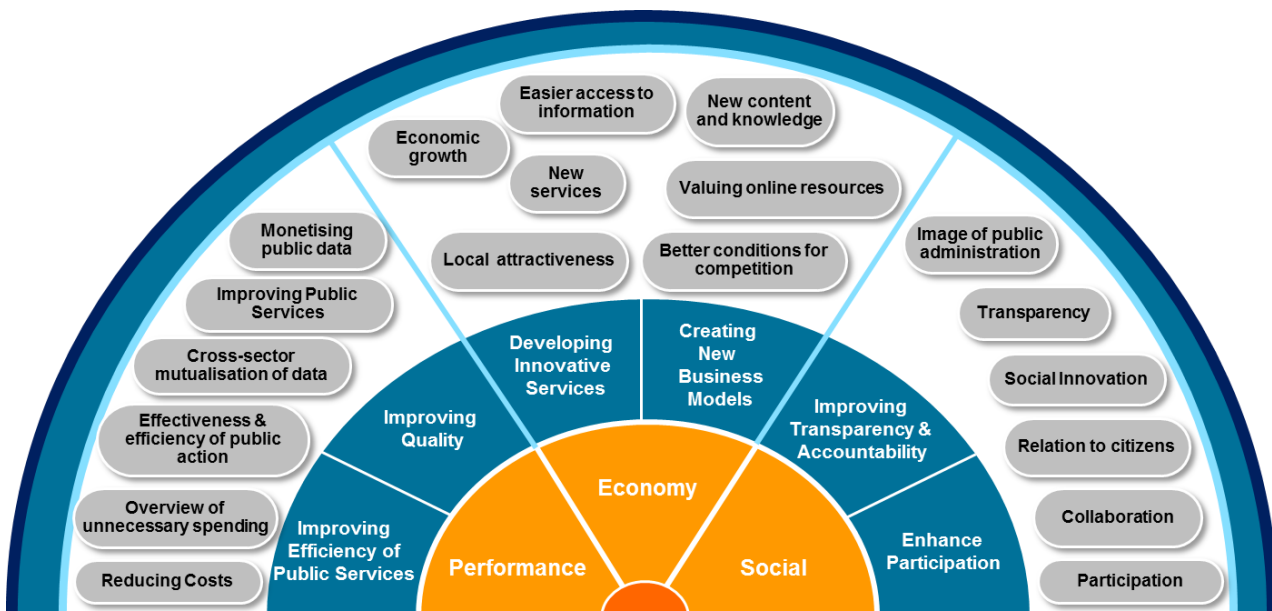

Figure 5: Benefits when Open Government Data is re-used

For more information about unlocking Open Data value, please also refer to the eLearning modules: *https://www.europeandataportal.eu/elearning/en/module2/#/id/co-01*.

Do you want to learn more about Open Data publication and re-use examples in Europe? Dive into more than 300 Open Data use cases: *https://www.europeandataportal.eu/en/using-data/use-cases*

EUROPEAN DATA PORTAL

If you want to learn more about Digital Transformation and Open Data, you can read the report: *https://www.europeandataportal.eu/sites/default/files/edp_analytical_report_n1_-_digital_transformation.pdf*

For a better understanding of how Open Data contributes to developing the smart city of the future, please refer to our reports on Open Data in the cities available here: *https://www.europeandataportal.eu/en/what-we-do/factsheets-and-reports*

A growing number of new initiatives and start-ups re-use Open Data. Several organisations generate case studies and keep track of start-ups that re-use Open Data. Please look at the box below for more links to examples and case studies of Open Data re-use. Furthermore, most (national) Open Data portals have a page with re-use examples of their data. It is worthwhile looking at those as well.

For more value stories, look at the national Open Data portals and their case studies. The following webpages are useful as well.

Value Stories by Open Knowledge:

http://opendatahandbook.org/value-stories/en/

Code for America success stories:

http://www.codeforamerica.org/

The Open Data Institute Case Study page:

http://theodi.org/case-studies

And the Open Data 500 website:

http://www.opendata500.com/

2. How to build an Open Data Strategy

Before starting to publish any Open Data, it is important to have a clear strategy in place that defines the key goals and sets the ambition. This chapter will address these key ingredients for a successful Open Data initiative as well as addressing barriers that one might face along the way and how these can be best tackled.

Figure 6: How to build an Open Data Strategy

2.1.　　Setting the ambition

Before taking action, you need to define what you want to achieve. This is often called "setting the ambition." Defining your ambition implies answering a series of questions such as: where do you want to stand? And by when? Will all data be available by default? Is all data stored centrally? In order to set a clear ambition, follow the steps as visualised below. Start by defining the As-Is situation, before going into the definition of the To-Be Situation and define measurable goals.

Figure 7: The steps for setting the ambition

2.1.1. Defining the As-Is Situation

In order to set the right ambition you first need to define a clear picture of the current situation. This we call the 'As-Is situation'. To get a clear picture of the As-Is situation in your organisation, the following 4 steps will guide you through the assessment of your current situation:

1. Gather a representative group
Identify key representatives of different sectors, units and departments of your organisation that can help you create the right As-Is situation.

2. Identify what the units do

Which units collect data? Which ones use data? Which ones produce data? What type of data do they gather? What format? A visualisation of the situation might help to create an overview.

3. Is data centrally organised?

The eventual effort required highly depends on two factors: the organisational data structure and the willingness of people. Throughout the organisation, data is generated by multiple units and is maintained at the same level. As a result, data is stored in numerous places with multiple responsible people dedicated to its storage and quality. On the other hand, data can also be centrally stored by a single data storage and maintenance unit, including a team delivering the service of data storage, maintenance and delivery. When your data is generated by multiple units and they all have their own policies and ways of working and people, you can imagine that implementing a new way of working becomes more challenging than when data is centrally managed.

4. Is data currently published?

If so, very good! This can be a good starting point. Find out what the current policies are around this process and try to identify best practices.

If no data is published yet, there is no need to worry. You can contribute to launching the process, which has its benefits as well and enables avoiding former restrictions, processes or wrongly manipulated data that might blur the view on the root of your data.

One of the main questions where many Open Data initiatives end is: if we were to publish data, what data should we open?

This question frequently leads to taking a step back and ends up in a discussion about whether to publish data or not. Sometimes it ends up in the decision not to publish data unless citizens request for it. Try to avoid such discussions and check whether you can identify some quick wins within your organisation that are part of your end-goal.
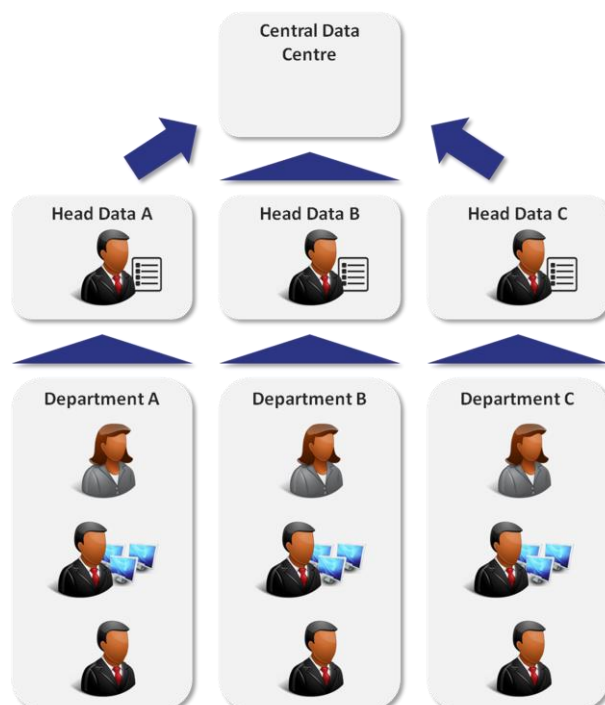
In summary, start by gathering a representative group, identify what the units do, check whether data is centrally organised and currently published.

Figure 8: Steps to define As-Is situation

Recommendation: Visualise the As-Is situation. This will help when you create an overview

This is a good practice example of how a data stream can be visualised.

2.1.2. Define the To-Be situation and define measurable goals

The next step is to detail your ambition. Therefore, you need to think of the To-Be situation: What do you want to achieve? Where will your organisation stand in 2 years' time? Or perhaps in 5 years' time? Discuss this with your group of representatives and create a clear picture of that To-Be situation. Again, we recommend you to create a visualisation of this situation.

Compare and define the goals

Put both visualisations next to each other and compare the As-Is and To-Be situation: What are the differences? What needs to be done? By means of defining clear and measurable goals, your organisation will be enabled to work towards those end goals and measure whether you have achieved them or not. While creating the goals, think of the primary reasons for publishing the data. These could be, for example, reaching the goal of becoming transparent or stimulating the economy. Make sure to be precise. Goals should be described in terms of scope, timing, deliverables and quantities among others.

Define measurable goals to demonstrate your success over time!

Do you want to know more about the current Open Data situation in the EU Member States? Read the Open Data Maturity in Europe report:

https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf

2.2.   Creating your strategy

The next step will be about defining how you will achieve these goals. The As-Is and To-Be situations are visualised and the goals have been defined. You are now entering the phase of creating a strategy. While creating the strategy, consider the questions you foresee and which actions you are going to take.

There should be a policy and a direction in place once the Open Data initiative is implemented. If not already done so when capturing the As-Is and To-Be situation, create a list of people within your organisation you will have to contact and involve. Create a scope, timeframe and planning, and start executing your project. Make sure your plan includes the topics as shown in the figure below.

As a summary, here are the nine topics every Open Data policy should cover:

1. **List of data (categories) within your organisation**
2. **Assessment of your organisation's data status, both current and future**
3. **Alignment with legal aspects**
4. **Actions that are needed to reach the goals**
5. **Technical implications: choices made or to be made**
6. **Budgetary implications**
7. **Team or Single Point of Contact with clear roles and responsibilities**
8. **Time plan to ensure the above is organised into time and can be monitored**
9. **Key Performance Indicators to measure your progress in realising your goals**

Figure 9: the 9 key topics every policy should cover

Poland, one of the Open Data "fast-trackers" in 2017 in terms of their Open Data maturity, created an Open Data strategy. It encompasses compliancy checks, a shared openness agenda and Open Data trainings. Overall, Europe scored high in the Open Data Landscaping of 2017, when it comes to having an Open Data policy in place with a performance of 79% in this indicator.

OPEN DATA PROGRAMME

Prepared by the Ministry of Digital Affairs
September 2016

https://danepubliczne.gov.pl/article/open-data-programme

EUROPEAN
DATA PORTAL

In reaching the To-Be state you might consider also set up an Open Data governance structure to ensure the involvement of all relevant (internal) stakeholders and when needed, the engagement of further external stakeholders.

The city of Vienna (Austria) adopted a comprehensive approach to their Open Data initiative that relies on 5 layers:
- Political layer (the Open Government Data was included in the 2010 programme of the city)
- Organisational layer (a virtual OGD competence centre was created which provided coordination and support for each department; each department was asked to appoint an OD officer)
- Legal layer (legislation, regulation and licensing norms were set up to accompany the process; the existing EU legislation and standards were implemented)
- Technical layer (the technical infrastructure for publication of both public sector information and geospatial data was ensured)
- Community and stakeholder management layer (both external and internal stakeholder management strategies were designed to ensure inclusiveness and dialogue)

A typical Best Practice from the Open Data in a Nutshell section

2.2.1. Publishing data: 'Open by default' or 'Closed unless'?

Concerning the discussion of what data to publish, organisations often end up choosing to close their data unless there is a clear purpose. This is a missed opportunity. Why? Because it is impossible to identify all opportunities different types of data create from the back of one's head; the opportunities are endless. You might classify a dataset as useless whereas a solutions architect has valuable ideas for the use of that set. Furthermore, assessing the usability of data case by case is a timely and unnecessary effort.

Be Open: others might see the value of a given data set where you might not

Therefore, we advise you to choose the 'Open by default' as a standard for your organisation, when it is legally feasible of course. However, if the assessment of the initial list of data shows that a very low percentage of data can be published, it might be wise to choose a case by case assessment.

2.3.   Drafting an Open Data Policy

Your Open Data policy is one of the most effective documents to detail your ambition and the way you intend to realise it. It will support your implementation and set the standard for the field. It will create the transition, increase the transparency of your organisation and ensure the best use of your data! The translation of your Open Data strategy into a solid policy is of great importance to ensure its successful implementation. First, an overview of the policy is given. Afterwards, the content of the policy is further detailed.

EUROPEAN
DATA PORTAL

2.3.1. The Open Data Policy: an overview

What is part of the policy? What data will you publish? Under what conditions? When? How often? And why? For what expected impact? What benefits? The policy will answer all the questions for the people within and outside your organisation. Take enough time to design your policy and make sure the decision makers within your organisation endorse it.

There are several policies available to look at. In line with your plan, make sure that the policy lands on the desk of those who are responsible for releasing data. You can expect many questions from stakeholders concerning the practicalities, such as budget, technical and practical aspects and legal boundaries. Therefore, make sure sufficient consultation and involvement takes place.

Your policy should cover topics such as definitions and expected benefits, the scope of the policy and expected outcome, legal aspects, etc. You may also consider defining data types and data quality and mandating an organisation, department or unit as responsible for the implementation of the Open Data policy.

Figure 10: Topics to include in your policy

2.3.2. The content of your policy

Definitions and benefits, scope and goals, legal aspects, data types and quality, point of contact are all important aspects to include in your data policy. These are all described below:
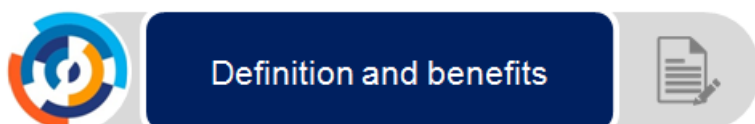
In the policy, it is important to indicate what data you want to publish, including a clear definition of Open Data. For example, part of the definition of Open Data could be that it is not only free to use, but also free of charge. Include definitions of the terms you use to clarify the scope. It is helpful to emphasize the benefits by including several examples that make explicit 'what's in it for them'. You can use examples from this Goldbook. Try to translate these benefits to specific benefits that fit the vision of your organisation. For example, having data available not only for externals, but also internally

EUROPEAN DATA PORTAL

for your colleagues, does that increase efficiency within your organisation? Then apply this argument as a key driver for convincing colleagues about your Open Data initiative.
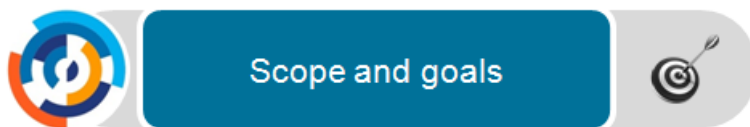
The scope is important to bring focus to the Open Data initiative. For example, it is possible to prioritize the release of data from specific departments first. Include what is part of the ambition, and what is not, by making your goals explicit and incorporating those in your policy. Describe the purpose, for instance, to improve transparency of your organisation and support the growth of the economy.
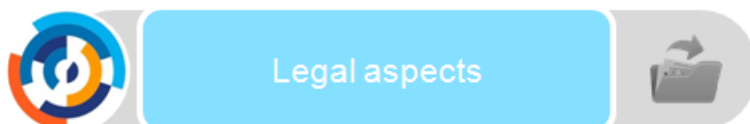
The implementation of Open Data has to be in line with current legislation. Publish datasets under an open licence. These legal implications are of major importance for any stakeholder dealing with Open Data. It is your responsibility as a publisher to be up to date with the most recent legislative rules applicable in your country. Your policy should address the following legislative topics.

Licensing
The data is not freely re-usable if you do not attach a licence to it. Thus, rule number one with regard to legally opening up your dataset is to attach an appropriate licence to it. There are many different types of licences you can apply, for instance one of the Creative Commons (CC) licences. A Public Domain Dedication is part of the CC licences and important to attach as it indicates that the public domain is the author of the data. More information is available on their website:

https://creativecommons.org/licences/publicdomain/

For more clarifications, the European Data Portal has a special Licencing Assistant:

https://www.europeandataportal.eu/en/content/show-license

In addition, the Open Data Institute created a comprehensive publishers guide to Open Data licensing:

https://theodi.org/guides/publishers-guide-open-data-licensing

Furthermore, the European Commission has published an introduction to data and metadata licensing:

https://joinup.ec.europa.eu/community/ods/document/tm25-data-metadata-licensing-en

Legislation: The Public Sector Information Directive
Every country has its own specific legal regulations when it comes to data. Therefore, all EEA[1] countries should have transposed the 2013/37/EC PSI Directive into the legislation of their country. The status of the PSI Directive transposition per country can be accessed here:

https://ec.europa.eu/digital-agenda/en/implementation-public-sector-information-directive-member-states

---

[1] European Economic Area

Intellectual Property

Intellectual property broadly includes everything created by the human mind. Open Data is free of intellectual property, as it is free to download, manipulate and re-use for any purpose, by anyone. It is possible to limit the re-use of Open Data by setting some legal boundaries to protect the provider, by adding restrictions of re-use in the licence.

Privacy

Data is sensitive and one must be cautious not to act in violation of privacy regulations. Make sure data is cleaned thoroughly and the appropriate level of anonymisation or pseudonymisation is applied to prevent the identification of a specific individual through your data. There are several legislative matters at EU level that determine how to handle data protection and how that applies to PSI. Re-use is only allowed in full compliance with the personal data protection rules. From May 2018 onwards, the General Data Protection Regulation (GDPR) will replace the Data Protection Directive, altering, in some key aspects, the laws surrounding the open publication of personal data. More information on the privacy component of Open Data can be found via the following link:

https://www.europeandataportal.eu/sites/default/files/open_data_and_privacy_v1_final_clean.pdf

Liability

Another important legal aspect is the liability of your organisation. Make sure all regulations are followed to prevent liability issues. The data that you publish should be reliable, which means with limited or no errors and anonymised.

Commercial law

Less likely to be a concern, but certainly a current debate: commercial law. Keep in mind that some datasets might intrude on the competitiveness of companies who have their business model built around particular data, which is now accessible by anyone.

Data is available in numerous types, formats and quality. Releasing relevant and quality data requires effort, but serves multiple purposes:

- It is easy to re-use data in Open Data formats
- Relevant data is used more frequently
- High quality (complete, well-documented) datasets enable re-users to identify its value for their purposes quickly

It is important to describe data types, topics and quality in your policy. It will serve as a standard for most of the datasets published by the public bodies and will increase the overall quality of your Open Data initiative and thus increase its effect.
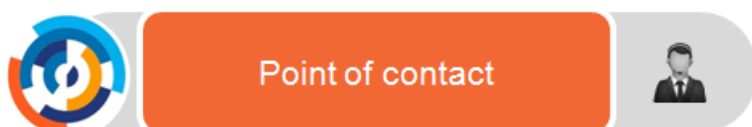
As a reference, it is useful to mention who is the responsible Point of Contact within your organisation. He or she will act as a single point of contact for queries related to Open Data within the organisation.

2.4. Overcoming barriers in publishing Open Data

Despite the wide variety of benefits, governments might still not acknowledge the potential of Open Data and seek for excuses not to publish the data as Open Data. This section offers a summary of frequently used excuses and an answer to the misconception.

The most frequent barriers recorded are

- Data is not interesting,
- The purpose or benefit for the organisation is unknown,
- There will be too many user requests for data,
- Users will draw superficial conclusions from the data,
- Data is not sufficiently accurate to be shared,
- It will cost too much to transform the data to a standard format,
- There is a risk to get a negative reputation,
- Publishing low quality data could harm the image of the public sector organisation.

The picture below offers a remedy to each barrier, see also the *Analytical report on barriers*:

Our Data is not interesting	Let others judge how interesting or useful it is. Even niche datasets have people that care about them. Remember the 'Open By Default' discussion on the previous page?
We don't see the purpose or benefit for our organization	Publishing PSI as Open Data brings numerous benefits, both internally in your organization, as well as externally to citizens and the private sector.
We will receive too many user requests on our data	On the contrary, when you provide raw data, end-users will extract and combine the data according to their needs. This will potentially reduce the number of requests. This is, for example, the case for CORDIS datasets which are accessible in bulk download files via the EU Open Data Portal.
Users will draw superficial conclusions from the data	In any form of publication the risk is there. Generally, the data is accompanied by supplementary documentation which put the data into context.
My data is not sufficiently accurate to be shared.	You cannot keep data always to yourself, there is a time of maturity; a time when data must be made available for reuse. Again, let others judge how interesting or useful your data is.
It will cost too much to put the data into a standard format.	When planned during the production phase, the production of or conversion to an open standard format requires little effort.
Am I taking a risk of getting a negative reputation?	One of the main objectives of the EU open data policy is that it leads to the wider use and to the spread of Union information, enhancing the image of openness and transparency of the institutions.
Publishing low quality data will harm our image	By means of user feedback you will have the opportunity to improve the quality of your data. Be open in your communication when publishing data.

Figure 11: Summary of barriers for publishing Open Data

2.4.1. Ensuring organisational alignment as a Key Success Factor

Ensuring organisational alignment for Open Data initiatives and implementing it in a sustainable way is essential to the success of an Open Data initiative. The Open Data Institute has performed a thorough analysis of change in organisations that publish their data as Open Data. They performed an extensive literature review regarding change management and tested their findings through

interviews with seven countries (Broad, Smith, Duhaney & Carolan, 2015). Their findings are summarised in a comprehensive article that elaborates on managing change in general, and how to successfully apply the basic principles to an Open Data initiative.

EUROPEAN DATA PORTAL

The 12 main principles of Open Data and managing changes:

	Principle	Description
	Articulate your vision	With clear examples of benefits Open Data will bring
	Secure support	For your Open Data initiative from both senior/political leadership and government officials within departments before launch
	Combine top-down leadership	For your Open Data initiative with support for individual or frontline champions in government
	Build open communication	And mechanisms for feedback into your Open Data initiative from the outset, both inside and outside government
	Set out some quick wins	For instance releasing a certain number of datasets, or supporting a pilot use-case – as part of a long-term goal for Open Data
	Be flexible and responsive	To the strengths and needs of different departments and teams
	Consolidate	Your change management efforts
	Ensure there are people	With responsibility for change management as part of your Open Data team
	Seek out and foster stories	Around the impact of Open Data
	Foster external support	Within industry, civil society and academia to drive continued demand for Open Data
	Introduce opportunities	For civil servants to take part in ongoing learning about Open Data
	Build metrics	To regularly evaluate Open Data

Figure 12: Summary of basic Open Data principles

2.5.   Critical success factors

The success of an Open Data initiative depends on three dimensions, namely:

1. **the quality of Open Data publication (e.g. accuracy, completeness, timeliness, and consistency)**
2. **the use of Open Data**
3. **documenting the emerging impacts and benefits.**

So far, little evidence exists regarding measurable factors that influence the success or failure of Open Data initiatives. The most critical success factors seem to be around addressing legislation, regulation, and licences. These aspects are very dependent on the local context. In addition, several success factors (e.g. sustainability of publication process, user feedback) appear to be applicable more universally.[2]

---

[2] Based on Open Data Disclosure and Use: Critical Factors From A Case Study by Anneke Zuiderwijk, Iryna Susha, Yannis Charalabidis, Peter Parycek, Marijn Janssen

EUROPEAN DATA PORTAL

An initiative is any activity that aims at improving the publication and/or use of Open Data, including initiatives on different levels (e.g. international, national, local) and by different stakeholders (e.g. civil servants, citizens, universities). Some factors are only critical for the success of data publication, while others are only critical for data use. It is therefore best to determine success factors regarding data publishing on the one hand and data re-use on the other.

Besides, the context of the initiative may also determine the criticality of a success factor. For example, multilingualism is critical when the Open Data initiative is organised globally and involves datasets from different countries, while this is less important for initiatives on a local or national level. Furthermore, whether the data is re-used for commercial or non-commercial purposes may influence the criticality of success factors for Open Data use. If there is a specific need to make Open Data use profitable, it is likely to lead to different success factors than in the case of not-for-profit data use.

2.5.1. Critical success factors for publishing

The most critical category of factors for the disclosure of Open Data seems to be **legislation**, **regulation**, and **licences**.

Furthermore, the sustainability of the Open Data initiative is a category of factors that is most critical for Open Data publication. Essential factors concern identifying the need for data, ensuring the continuity of data supply (including timely and automatic updates of data), and being transparent towards Open Data users about the conditions under which data publication takes place. Factors regarding accessibility, interoperability and standards that were critical for Open Data publication success are multilingual metadata and data, the use of standards (for data, metadata, licences, URIs and exchange protocols), the integration of metadata

schemas and federated controlled vocabularies, the provision of various types of metadata in line with metadata standards, and the supply of APIs for Open Data provision in the form of service needs.

Finally, within the category Open Data platforms, tools and services, the critical factors are the presence of one central portal, which combines data from many different governmental organisations, the integration of frameworks for assessing data quality and usability of data and platform, providing continuous feedback to developers and administrations and the development of a clear user interface. Stewardship and the development of a management plan seem critical success factors. Therefore, it is important to start with a clear Open Data strategy.

2.5.2. Critical success factors for re-use

Legislation, regulation, and licences also seem to be critical for the re-use of Open Data. For example, the provision of information on the meanings and implications of licences, and on privacy legislation and how Open Data can be used in compliance with this legislation, are critical. Furthermore, success stories are critical, especially the provision of readily available examples of Open Data use (e.g. applications) to non-experts. Success stories attract a large user base. In addition, all factors related to feedback and sustainability are critical for the use of Open Data. The provision of mechanisms for governmental agencies to know how their data is re-used is important. Furthermore, to know what can be learned from the re-use of their data and how the publication of their data can be improved. Open Data users should know precisely the methodology of how the data was produced described in a scientific manner.

3. Technical preparation and implementation

From a technical point of view, publishing data can have a large impact. Publishing data involves several processes. In short, it involves collecting, preparing, publishing and maintaining data. In this chapter, we will highlight the most important aspects to keep in mind, namely Data management, Extracting, transforming and publishing data, Channels, Search and Pre-requisites, choices and accountability.

Figure 13: The technical implementation process

3.1.   Data management

Before publishing data as Open Data, your organisation should have an overview of your current data management structure. In the As-Is Situation visualisation, you can see how your data management is organised. It is wise to create a structured data management process first before publishing data. This section gives a brief overview of the various states of data management. Try to translate this information to your own As-Is situation and apply them to the strategy. You can identify the situation of your own organisation and from there on decide on which steps to take.

> **Recommendation: First, draft an overview of the data management practices within your organisation and make sure that this is clear and documented before publishing data**

In some cases, neither a specific structure nor governance has been set up. If the situation within the organisation is unstructured without governance, start the Open Data publication process by creating a data management infrastructure first.

3.1.1. Data management per instance

Data can be scattered over units or persons that all have the responsibility to manage their own data. This situation is an example of short-term data management and only recommended for organisations that occasionally deal with very small amounts of data (e.g. maximum of 10 sets that are updated maximum 4 times per year). For these smaller amounts of data, it remains manageable to create consistency over the data. For larger amounts, it is recommended to move towards a higher level of data management and introduce Master Data Management (MDM). See the picture below to see what that looks like.
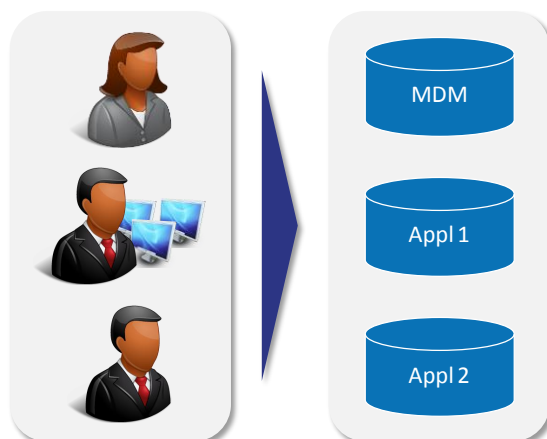
Figure 14: Data management per instance

3.1.2. Federated Data Management with a Centre of Expertise (CoE)

In this case, data management is still scattered over a variety of units, but a Centre of Expertise (CoE) is in place that is responsible for the management of data. Thus, in this situation there is more coordination and therefore, increased possibilities to create overall consistency. This state is more integrated and useful for organisations dealing with larger amounts of datasets. It is a first step towards fully integrated Master Data Management governance. However, in this case the CoE has no authority in enforcing consistency and therefore, has a more facilitating role.
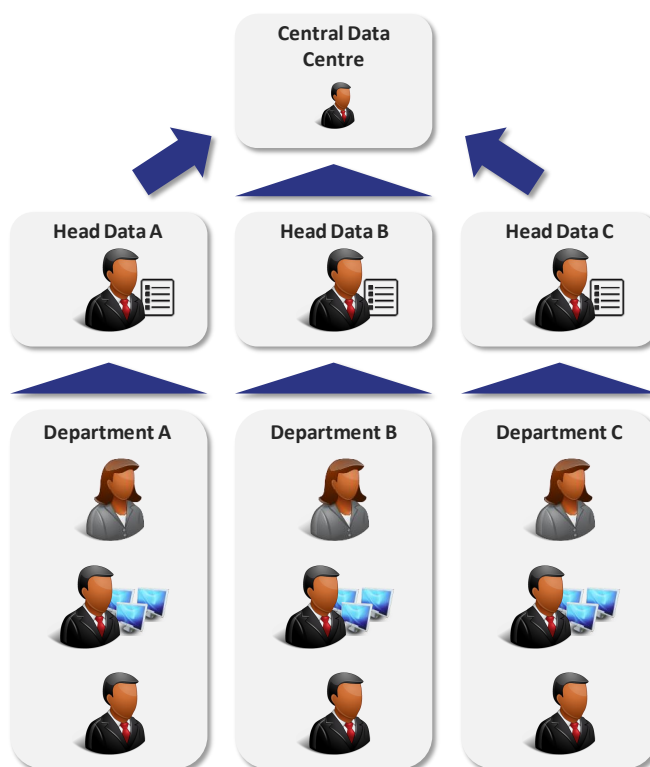
Figure 15: MDM with Centre of Expertise

3.1.3. Fully Centralised Master Data Management

This is the ideal state for organisations dealing with larger amounts of datasets. Certainly when the velocity of the data is high, central master data management governance is highly recommended. In this case, standardisation is determined at the highest level and consistency is created throughout the organisation. This will ensure smooth running processes and cost savings on the long run. Furthermore, this governance structure is most sustainable even when the organisation structure changes over time.

Figure 16: Fully centralised MDM

3.1.4. Implementing Master Data Management

Moving towards a suitable data management structure might be an impactful operation. If the organisation requires a change in data management structure, it is recommended to gain advice and expertise from external parties that have experience in data management change programs. **"Appendix 2 - Master Data Management Change Plan"** provides a detailed 10-step guide for data management change.

3.2.   Extract, transform, and publish

Publishing data as Open Data from a data source is a process that overlaps with data warehousing processes where data is extracted, transformed and loaded (ETL-process). In the case of publishing Open Data, the load phase is replaced by the publish phase. Depending on the ambition of the organisation, multiple scenarios are possible for implementing the Extract, Transform, Publish (ETP-) process.

The ETP-Process is the technical specification of how data flows through the organisation, transforms into a publishable dataset, and eventually is made public. The Open Data Lifecycle (explained in the following chapter) is part of the ETP-Process.
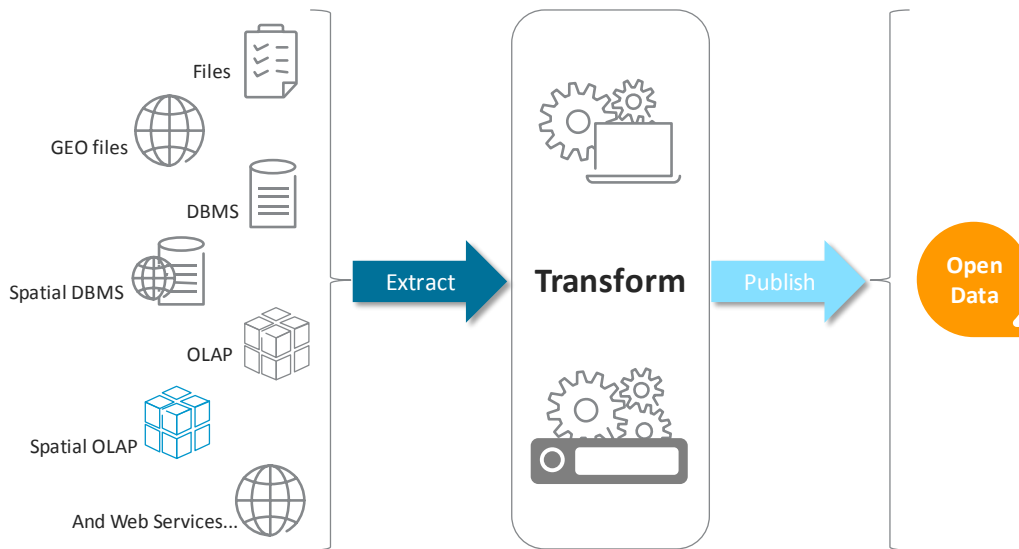
Figure 17: A visualisation of the Extract, Transform, Publish process

Data can be published from:

- An existing publication
- An existing database
- A source database
- An existing source system or package
- Different sources – consolidating data

Please take a close look at the suggested options in **"Appendix 3 - The Extract Transform Publish (ETP) Process"** and choose an ETP-process that is suitable for your situation.

It is important to keep in mind that once all technological aspects are in place, the ETP-process will be the central and recurring technological process for every single dataset that is published. Within the ETP-process, quality measures take place. These quality specifications for the ETP-process are determined in the Open Data Lifecycle, as discussed later.

3.3. Channels

Depending on the ambition and the strategy set, there are technical decisions to make. This section shortly elaborates on the channel options and different software solutions. Web download, data portal and API will be discussed. The technical implications of these channels are further presented in the following chapter on the Data Management Lifecycle.
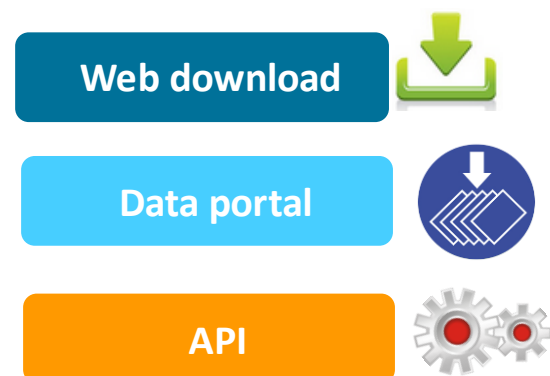
Figure 18: Different publishing channels

3.3.1. Web download

This is the simplest way to publish Open Data. The data can be published on the regular organisation's website by means of adding a separate section to the website that contains links to the datasets. This is the best way to initiate Open Data publication and is used often in situations where little amounts of datasets are to be published.

The software solution used could be any Content Management System (CMS), such as Drupal, Wordpress, Django, (possibly) in combination with an internal database.

3.3.2. Data portal

A more advanced way of publishing datasets is through a data portal, which is particularly useful if the publishing organisation has large amounts of datasets to publish that need to be updated regularly. The setup of a portal includes the installation of catalogue software (free, open source software is available) in order to make a structured set-up of the data. Portals are web-based and thus require a website.

Another approach is to publish on someone else's portal. This is a low-cost solution and is very applicable in some situations. By adding your data to existing data portals, the discovery of your data will increase. Either these portals need your data to be uploaded to their sites or the portal crawls the data from your website automatically. Adding your data to harvesting portals, such as the European Data Portal, sometimes requires additional settings, such as metadata formats, use of vocabularies and other specifications.

Check the data portals for additional information. Do not hesitate to contact similar administrations or nearby regions to initiate a portal together.

3.3.3. API

An API (Application Programming Interface) is a more advanced and technical way of connection distinct systems or applications to one another. In the context of Open Data, an API is software that provides a system with direct access to data from another system. The API grants access to the use of the data catalogues and its functionalities. Without interference of any end-user interfaces such as portals or webpages, a third-party application can load the data by means of a request protocol. In short, it is particularly useful if the data should be up to date, directly accessed and re-used by third parties, and the application using the data needs direct access to the data in the database without any interference. For example, most mobile applications make use of API's for the retrieval of data.

Technical solutions such as CKAN and DKAN come with APIs that you can use.

3.4.   Search

3.4.1. Basic search

You should consider implementing a basic search functionality if you have a large amount of data on your website. Such search functionalities are built into the Content Management System (CMS) used for the website of the public sector organisation. Furthermore, a multitude of open source and

proprietary (commercial) search tools are readily available for use.

3.4.2. SPARQL

Next to basic search functionalities, there is the option of using SPARQL. This is a query language suitable for executing search queries for Linked Data. For technical re-users this is a great advantage. We recommend not to use SPARQL as a standard search option, but rather as an optional advanced search tool, as many re-users are not familiar with such technical query language. The European Data Portal for example also provides advanced users with the possibility of performing a SPARQL-search for the desired data.


```
PREFIX foaf:  <http://xmlns.com/foaf/0.1/>
PREFIX card: <http://www.w3.org/People/Berners-Lee/card#>
SELECT ?homepage
FROM <http://www.w3.org/People/Berners-Lee/card>
WHERE {
    card:i foaf:knows ?known .
    ?known foaf:homepage ?homepage .
}
```

Figure 19: European Data Portal SPARQL functionality and a SPARQL query example (Feigenbaum, L., 2015)

Recommendation: If you publish your data as Linked Data, include SPARQL as a search tool option to allow advanced users that are familiar with such query language to use it.

3.5.   Pre-requisites, choices and accountability

Once the data management structure, ETP-process and channel have been agreed upon, the following choices must be made and a responsible person needs to be appointed (Van Herreweghe, 2015).

- **Choosing the domain:** the public sector organisation may opt to use its own website for making the data available or it can do so via a separate website with its own domain name.
- **Choosing the hosting**: the public sector organisation must determine whether the data will be stored and made available on its own servers or whether it will use third-party servers for this.
- **Choosing the functionalities**: examine which database will be used, for instance whether a forum or a payment module is required. An assessment of the required server space, the total data

consumption and the required speed is needed.

- **Managing the website and/or the portal**: someone should be appointed to manage the website or portal. When doing so, it should also be determined which degree of availability is to be guaranteed and which level of website monitoring and security is needed.
- **Maintaining the services**: if data is made available through services, someone should be appointed to monitor and guarantee the availability, functionality and performance of the services.

4. Putting in place an Open Data lifecycle

Publishing any type of data is a process that consists of various sub-processes: *Collecting, Preparing, Publishing* and *Maintaining.* Applying this process will result in a structured Open Data system within your organisation. Look at the sub-processes and think of how you could implement this in your organisation. The upcoming sections will also explain key (technical) concepts of Open Data that you should be aware of in this context. Do not forget to include this process in your policy!

Figure 20: The Open Data Lifecycle

4.1.   Collecting data

With all preconditions in place, the data can be collected. Where should you start? What is relevant? This chapter goes into the details of collecting and identifying data. Collecting data can be approached from two angles: quick wins and thorough data management. It highly depends on the infrastructural choices within your organisation. Look at your strategy: Where will the data be managed? Will it be done centrally or is it processed at multiple units?

4.1.1. General collection process

Create a process for collecting data that suits your situation. The following is a brief description of steps that might come in handy while creating your collection process. This process includes mapping the currently available datasets, prioritizing the datasets, practicing, topics to publish and publishing categories.

Figure 21: Different steps of the collection process

Map the currently available datasets

Start your Open Data initiative by creating an overview of the data that is already available in your organisation. This is a quick win: the data is there and you will have a list of all data and where it is managed. Ask your data-managing colleagues to help you with this.

Prioritise the datasets

Not all datasets are relevant to publish right away. To prioritise your list, look at the following criteria:

- Can it be published (legally, politically, and organisationally)?
- Is it of the right quality (and thus does not need thorough manipulation before publication)?
- What about cleaning, anonymising, good quality and format?
- Does it belong to one of the high-value topics?

The datasets that meet these requirements should be prioritised: these are your quick wins. With this list, you have a complete overview of the data and you have identified what can be published, what not and what should be published first. Later on, you can choose to prioritise by demand or other parameters.

> **Recommendation: Create Quick Wins and start with those. Practice your collection process first to get acquainted with it. You will be able to improve it, and answer questions that are asked about it.**

Practice

Go through the collection process. What steps did you take? Who is responsible for the next part of the process? What is the standard process of collecting and prioritizing data? What will happen if new data is created or a dataset is updated? Learn by doing and document the steps.

The Irish Best Practice Handbook described a best practice around auditing your existing data, and suggests how to become aware of the datasets that are currently available within the organisation. See the Best Practice Statement below.

1. Each public body should carry out a data audit of the data they currently manage
2. Information on each dataset should be recorded according to the standard metadata format of the national Open Data portal. Information for each dataset should include:
    a) Potential for release as Open Data (governed by an 'Open by Default' principle)
    b) Legal information
    c) Organisational information
    d) Technical information
    e) Value assessment:
        1. Datasets recognised as 'high-value' datasets should be released proactively
        2. Data audit results should be made available on the national Open Data portal to enable users to request the publication (demand-driven publication)

EUROPEAN DATA PORTAL

Types of data to publish: the G8 Open Data Charter and European Commission guidelines

Data is created, stored, and distributed covering a large variety of topics and categories. However, not all types of data are of equal relevance. In 2013, the G8 came together to discuss governmental transparency, innovation and accountability. This discussion led to the creation of the "G8 Open Data Charter" (Cabinet Office, 2013): a summary of visions and principles for creating a transparent Government, the opening up of data and its quality and quantity.

Part of this charter holds valuable and useful guidelines concerning topics, data types and formats, and quality. The most relevant and high-quality topics are summarised in the following 14 categories:

Data Category (Alphabetical order)	Example of Datasets
Companies	Company/business register
Crime and Justice	Crime statistics, safety
Earth observation	Meteorological/weather, agriculture, forestry, fishing, and hunting
Education	List of schools; performance of schools, digital skills
Energy and Environment	Pollution levels, energy consumption
Finance and contracts	Transaction spend, contracts let, call for tender, future tenders, local budget, national budget (planned and spent)
Geospatial	Topography, postcodes, national maps, local maps
Global Development	Aid, food security, extractives, land
Government Accountability and Democracy	Government contact points, election results, legislation and statutes, salaries (pay scales), hospitality/gifts
Health	Prescription data, performance data
Science and Research	Genome data, research and educational activity, experiment results
Statistics	National Statistics, Census, infrastructure, wealth, skills
Social mobility and welfare	Housing, health insurance and unemployment benefits
Transport and Infrastructure	Public transport timetables, access points broadband penetration

Table 1: The G8 High Value categories of data

The purpose of this list of categories is to ensure that Data Holders focus on the release of the right and most relevant types of data. This does not mean that other categories of data cannot be published. The list above gives an indication of the topics that should have the highest priority, as these datasets are indicated as datasets with the highest potential value.

Based on the G8 Open Data Charter, and different European Commission studies, the European Commission published its own standards. These Guidelines on recommended standard licences, datasets and charging for the reuse of documents help to harmonise European efforts related to licences and datasets and accommodate the PSI Directive as a result.

Publishing categories

Next to gathering categories, there are publishing categories. You might want to publish your data under another set of categories than the G8 list. Other portals have created their own set of categories as well. Think of your data: under which categories are you going to publish your data?

To provide you with an idea of how to categorise your data, here is an example. Please look at the categorisations as a re-user. Try to imagine that you are looking for a single file: how will you navigate towards it? There are pros and cons of both large and little amounts of categories. Try to find out what suits your purpose best and what you, imagining being a re-user, prefer as a logical structure. The one requirement is that it is automated through metadata. Figure 22 below shows an example of the categorisation used by the European Data Portal linked to the DCAT Application Profile detailed in the next sections of this chapter.

Figure 22: Example from http://www.europeandataportal.eu/

Preparing data

Now that the data has been collected, it should be prepared for publication. As raw data is not useful for publication, a set of actions will prepare the data to be published. This section will provide more information about getting the data ready. Preparation can be done touching 4 topics: Quality, Technical Openness, Legal Openness and Metadata.
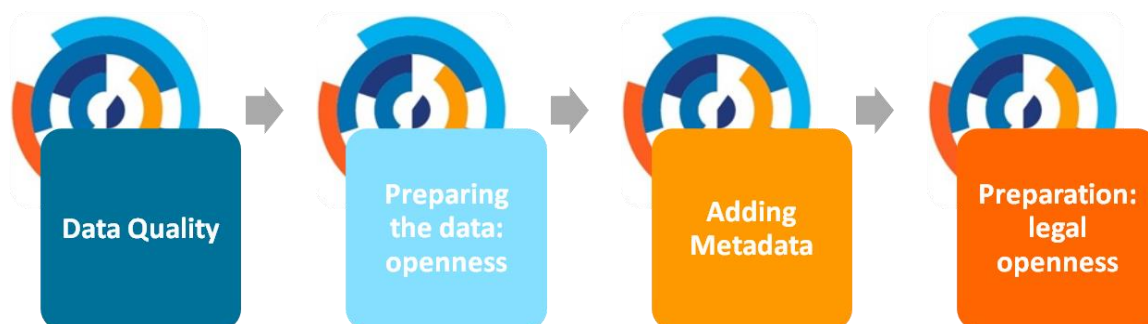
Figure 23: Steps to prepare data

EUROPEAN
DATA PORTAL

4.1.2. Data quality

This chapter highlights the following aspects of data quality: content quality, timeliness, and consistency.

Figure 24: Aspects of data quality

Content quality

Usefulness can be determined by the quality. The quality of Open Data, next to its discoverability, is one of the largest influencers of the success of Open Data. Quality concerns many aspects. This chapter covers the completeness, cleanness and accuracy of data.

Is the data complete?

Is your dataset complete? Completeness concerns various aspects. Every dataset should:

- Contain a header row with a single description of what is shown. This means that once a dataset structure is in place, it should not change when sources are added. In the metadata, the header should be described
- Be labelled with a version number. Once an update is done the dataset should get a new version number in order for the audience to keep track of changes
- Contain information about its origin. What is the data about, where does it come from and for what purpose has it been published?
- Be given a status: Draft, validated, final

Is the data clean?

Is your dataset clean? Cleanness concerns various aspects. Check the following aspects:

- Empty fields
- Dummy data and default values: are they correct?
- Wrong values
- Double entries
- Privacy sensitive information

Always check your data on these points and make sure that your dataset does not violate any of the legal constraints mentioned in "Legal Openness".

Is the data accurate?

Is your dataset accurate? Accuracy concerns various aspects. The most important aspects regarding accuracy:

- Is the data accurate enough for its potential purpose?
- Does its accuracy affect its reliability?
- Are the choices concerning interval described
- Does the data need aggregation or disaggregation?

Is the data accurate enough for its potential purpose?

Does its accuracy affect its reliability?

Are the choices concerning interval described?

Does the data need aggregation or disaggregation?

Figure 25: Questions regarding accuracy

Timeliness

Data changes over time. Historical data will remain stable, but recent data will be updated over time. Therefore, it is important to check data with regard to its timeliness regularly. For consistency purposes, it is wise to create an update process that keeps the data up-to-date. Be sure that the data contains a notion of its timeliness. This topic is closely related to the maintenance of datasets.

Consistency

Reading through the quality aspects of data, the consistency of the presentation of your data is of major importance. Imagine re-users correlating data from various sources, but all datasets differ in accuracy, use of terms and timeframe. As an example, if you change the field names of the data collected for managing waste each year, the data cannot be compiled from one year to the next. This makes it difficult to use datasets: it will require a large effort of manipulation. Therefore, make sure you use the standards and be consistent in publishing datasets of equal quality.

4.1.3. Preparing the data: technical openness

Data has been prepared in terms of quality. In this chapter, several concepts will be introduced: Linked Data, metadata and the 5-Star Open Data Model. To understand the 5-Star Model, you will have to understand the basics of Linked Data.
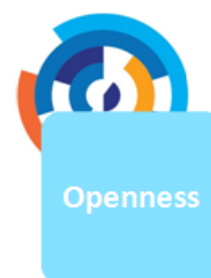
4.1.3.1. Linked Data

The concept of Linked Data increases the interoperability and discoverability of datasets. Linked Data is not the same as Open Data. Whereas Open Data concerns the openness of the data itself, Linked Data is a way of publishing Open Data as Linked Data or enriching datasets with Linked metadata. This is where it gets more technical. The definition of Linked Data:

"Linked Data is a set of design principles for sharing machine-readable data on the web to be used by public administrations, business and citizens" (Berners-Lee, 2013)

Linked Data are pieces of information that are linked through a graph connection. Opposed to other

relational descriptions of data, in Linked Data, a machine can walk through the graph and understand the content. This is seen as a revolution in the area of data storage and sharing: a computer can, to some extent, qualitatively interpret the data. This is possible, because the data is enriched with uniform descriptors. By means of these descriptors, the data is no longer a set of static content, but is described and can therefore be interpreted, regardless of any distinguishing factor such as language or file type.

We will provide you with a comprehensible example from the Educational Curriculum for the usage of Linked Data (EUCLID) module 1 (EUCLID, 2014) and we will explain the basic concepts attached to Linked Data through this example.

Ontologies

Datasets usually encode facts about individual objects and events, such as the following two facts about the Beatles (shown here in English rather than a database format):

The Beatles are a music group
The Beatles are a group

There is something odd about this pair of facts: having said that the Beatles are a music group, why must we add the more generic fact that they are a group? Must we list these two facts for all music groups, not to mention all groups of acrobats or actors, etc.? Must we also add all other consequences of being a music group, such as performing music and playing musical instruments?

Ontologies allow more efficient use of data by encoding generic facts about classes (or types of object), such as the following:

Every music group is a group
Every theatre group is a group

It is now sufficient to state that the Beatles (and the Rolling Stones, etc.) are music groups, and the more general fact that they are groups can be derived through inference. Ontologies thus enhance the value of data by allowing a computer application to automatically infer many essential facts that may be obvious to a person, but not to a program.

Linked Data make use of several techniques, among which **RDF**, vocabularies and **URIs**. Many data catalogues and Open Data portals that aim to publish Linked Open Data use predetermined vocabularies in order to remain uniform. The European Union Catalogue Specification is called **DCAT-AP**. Therefore, it is recommended to use DCAT Application Profile. A brief description of these terms is presented below.

RDF

The RDF Framework (Resource Description Framework) is the basic principle of Linked Data. It is the new general syntax for representing data on the web. This syntax is a link (URI – Unique Resource Identifier) that is built from 3 descriptors, which all together are called a Triple. By describing an object with this triple, it becomes linked. As terms can differ from sentiment, the structured way of describing them through the RDF triple overcomes this. Furthermore, as many terms are described through RDF terms, they can be linked to each other.

Every piece of information expressed in RDF is represented as a **triple**:
- **Subject** – a resource, which may be identified with a URI.
- **Predicate** – a URI-identified reused specification of the relationship.
- **Object** – a resource or literal to which the subject is related.

Example: name of a legal entity:

http://example.com/org/2172798119 has as legal name "Nikè".

Subject   Predicate   Object
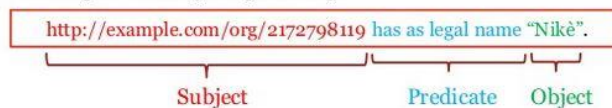
Figure 26: The idea of an RDF Triple

A basic introduction to RDF: *http://www.linkeddatatools.com/introducing-rdf*

RDFa

The frequently used technique of RDF on data portals is RDFa: embedding RDF in HTML. A comprehensible quick presentation on RDFa:

http://www.slideshare.net/fabien_gandon/rdfa-in-a-nutshell-v1

Always publish your metadata embedded in HTML with RDFa. An example:

```
<html>
<head> ... </head>
<body>
...
<div resource="http://example.com/org/2172798119"
typeof= "http://www.w3.org/TR/vocab-regorg/RegisteredOrganization">
<p>
<span property=" http://www.w3.org/TR/vocab-regorg/legalName">Nike<span>
Address: <span property="http://www.w3.org/ns/locn#fullAddress"> Dahliastraat
24, 2160 Wommelgem </span>
</p></div>
</body>
```

Figure 27: *An RDFa embedded in HTML example*

URI

The term URI stands for Unique Resource Identifier and can refer to text, Uniform Resource Name (URN) or Uniform Resource Locator (URL). Its main function is to identify something. In general, in the case of Linked Data, URIs are triples in the form of a URL (*http://www.europeandataportal.eu/*) or vocabulary specific identifiers. Detailed information about URIs:

https://joinup.ec.europa.eu/sites/default/files/c0/7d/10/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf

Additional information

A deep-dive presentation about Linked Data:

http://europeandataportal.eu/elearning/en/module13/#/id/co-01

Figure 28: European Data Portal eLearning Module on Linked Data

4.1.3.2. Metadata

It is important to ensure that your data can be found. The term usually applied to this is the *discoverability* of data. Essential for discoverability is metadata. Metadata describes the dataset itself (e.g. date of creation, title, content, author, type, size). This information about the data needs to be added to the catalogues to help discover the data. Metadata needs to be both human understandable and machine readable. If it is published as Linked Data, the discoverability of the data is greatly increased. Metadata does not only serve the purposes of description and discovery, but also renders itself as essential for the scope of contextualisation (relevance, quality, restrictions (rights, costs)), as well as for matching users and software to data available on the internet.

Metadata has a large influence on the re-use of Open Data. It will increase the discoverability and the re-use of your data. Therefore, take the time to inform the re-user about the quality of the dataset by providing rich metadata. This will make the usability of the dataset better. Metadata has been defined by the W3C Foundation as (W3C Foundation, 2015):

"Metadata is structured information that describes, explains, locates or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called "data about data".

In a nutshell, Metadata helps:

Discover data

Understand the nature and structure of the data

Find te origin of the data and under what terms it may be used
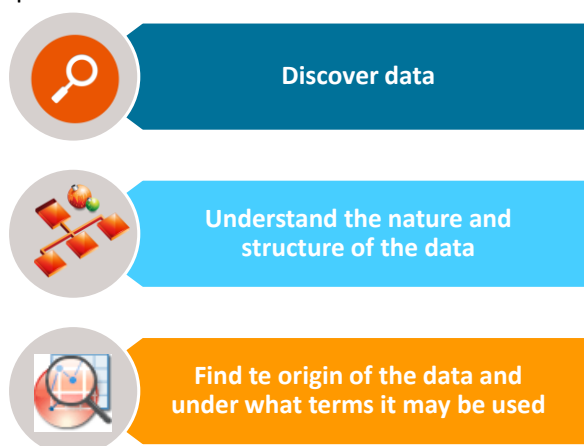
Figure 29: Important reasons to add Metadata

Recommendation: For a full description of best practices with regard to metadata, please go to the website of the W3C Foundation: http://www.w3.org/TR/dwbp/#metadata

Here is an example of the metadata that would be used to describe the Beatles:

```
@base <http://musicbrainz.org/>.
@prefix mo:<http://purl.org/ontology/mo/>.
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#>.
@prefix owl:<http://www.w3.org/2002/07/owl#>.
@prefix dbpedia:<http://dbpedia.org/resouce/>.
@prefix bbc:<http://www.bbc.co.uk/music/artists/>.


<artist/b10bbbfc-cf9e-42e0-be17-e2c3e1d2600d#_>
        rdfs:label "The Beatles";
        owl:sameAs dbpedia:The_Beatles,
```

Figure 30: Describing the Beatles as metadata

Datasets can be enriched by descriptions, making the interpretation easier. Metadata within the context of Linked Data has even more value: by enriching metadata with URIs, the data can be linked. This enhances the discoverability and interoperability of data incredibly. If you publish metadata with your data, it is recommended to enrich your metadata with URIs. Important to know is that metadata is a necessity if you want to be harvested by data portals such as the European Data Portal.

> **Recommendation: Always publish your metadata as Linked Data. This increases the discoverability and interoperability of your datasets**

Metadata standards

Metadata provides information on the data and hence enables access to digital objects. Standards rend themselves essential in the metadata context as they provide a common structure for the "data on the data", and help ensure the interoperability of systems and enhance the discovery of and access to data. Here are the most known metadata standards used in the Open Data context:

- Common European Research Information Format (CERIF) – EU recommendation to Member States and it is used for all research information.
- Dublin Core (DC) – used to describe webpages.
- Data Category Vocabulary (DCAT) – used for datasets on the internet, also based on DC.
- Comprehensive Knowledge Archive Network (CKAN) – used in government Open Data sites.
- eGov metadata Standard (eGMS) – based on DC.
- Inspire – used for datasets with geospatial coordinates.

Metadata Best Practices

Providing qualitative metadata is a complex but necessary practice. The W3C foundation has developed guidelines and best practices to support data holders. Furthermore, interoperability with the European Data Portal is crucial. This avoids costly crosswalks and mappings between datasets. Hence, the use of the DCAT-AP is strongly encouraged. To summarise, publish the metadata with the data using a machine-readable format and standard terms to define the metadata. In addition, describe the overall features of the dataset with information about local parameters, licence, origin and quality.

EUROPEAN
DATA PORTAL

	Publish the metadata together with the data
	Use machine-readable formats to provide metadata
	Use standard terms to define metadata
	Describe the overall features of a dataset
	Provide information about local parameters (date, time, value types, language)
	Provide license information Provide information about its origin
	Provide information about the quality

Figure 31: Summary of metadata best practices

Publishing metadata

Although a metadata-set is closely related to the dataset it describes by providing helpful data about this particular dataset, it can sometimes be useful to provide the metadata at multiple places. This enables different Open Data portals to address different audiences. For example, the European Data Portal thrives to be a place where all Open Data portals of the EU Member States share the metadata of their Open Data. In this way, citizens and businesses will have one single place to access metadata data or in other words, access the data about the data available from all over Europe. The participating countries' own Open Data portals will stay active as portals hosting the data and responsible for their respective regions, enabling users with specific needs to use a fitting Open Data portal. An even more fine-grained approach can be seen when looking at small Open Data portals maintained by cities or other administrative areas.

Thus, you will often find sets of metadata in different Open Data portals describing the same underlying dataset. Sometimes, these are copies from the portal the data was originally provided on, other times less metadata is provided when the metadata schema applied does not allow the display of the full metadata. The process of providing the metadata to different portals can either be done technically by the Open Data portal itself (called harvesting, explained next) or should be done manually.

Figure 32: Steps to publish metadata

Harvesting metadata

Large Open Data portals often act as an aggregator of smaller Open Data portals. They regularly check for new data in smaller Open Data portals and copy the metadata found so that users will find it there as well. This process is called harvesting. Open Data portals usually do this in the background without

the user noticing it. Open Data portals often provide an API with which they provide their data or metadata in a machine-readable format. These APIs can be used by other Open Data portals – or any other user – to read the data and copy it into their own database. Sometimes the data has to be transformed to a different format, because a different categorisation is used.

Depending on the API protocol, the harvesting entity can apply filters if, for example, only a subset of the data from the harvested portal is desired. By using harvesting, portals will have a greater database and can address a bigger or more specific audience without having to rely on users providing the metadata manually.

> **As an example, see the harvesting criteria of the European Data Portal**
> **https://www.europeandataportal.eu/en/providing-data/how-to-be-harvested-by-us**

Mapping the metadata

For easy metadata inclusion, map the metadata. To do so, use the standard Linked Open Data vocabularies (DCAT-AP) to create a table of properties and URIs to enable easy adding of the metadata to the file. Make the distinction between metadata with regard to the dataset itself (title, description, licence), and metadata with regard to the distribution (URL, format, status).

Controlled vocabularies

The European Commission has created a Linked Open Data specification for metadata called DCAT-AP. This increases interoperability between all data portals in Europe. For instance, when talking about file formats, there are various standards.

- DCAT, Data Catalogue Vocabulary
  http://www.w3c.org/TR/vocab-dcat/
- DCAT Application Profile is not a vocabulary, but a specification for metadata descriptions of EU governmental data and portals

Please look at the list of the most recent DCAT-AP publications to learn more about controlled vocabularies, also including information on extensions (GeoDCAT-AP, StatDCAT-AP):

https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe

For more general information and training about metadata, please look at the following links:

http://w3c.github.io/dwbp/bp.html#metadata

and

https://www.europeandataportal.eu/en/resources/elearning

and

https://theodi.org/guides/marking-up-your-dataset-with-dcat

Multilingual Thesauri

For international interoperability, it is useful to make use of multilingual thesauri. This means that you use a standard vocabulary set of words, which can be translated to other languages more easily. Eurovoc is a multilingual thesaurus. Please see: *http://eurovoc.europa.eu/*

Publishing the metadata

Most portal software solutions come with integrated metadata creation modules. In this case, metadata is created by filling in predetermined fields or by choosing from drop-down lists.

> **Recommendation: Always publish your metadata as Linked Data. This increases the discovery and interoperability of your datasets**

4.1.3.3. The 5-Star Open Data model

Publishing high-quality Open Data requires some effort. The W3C Foundation has created a basic model for Open Data with regard to quality: the 5-Star Open Data model. The 5 stages of Open Data are:

★	Make your stuff available on the web (whatever format) under an open licence
★★	Make it available as structured data (e.g. Excel instead of image scan of a table)
★★★	Use non-proprietary formats (e.g. CSV instead of Excel)
★★★★	Use URIs to denote things, so that people can point at your stuff
★★★★★	Link your data to other data to provide context

Table 2: Descriptions of all stages of the 5-star Open Data Model

The 1-Star Stage: Publishing your data

Stage 1 of the 5-Star Open Data model can be achieved by publishing your data. This can be done in various ways, via download, bulk download, or API's.

The 2-Star Stage: Making it available as structured data

The power of Open Data lies in its re-usability and stimulates interoperability of systems and services. Data formats can be clustered into 2 categories:

- Structured data (machine- and human-readable)
- Unstructured data (or human-readable)

Structured data is developed to be **processed** by machines and is thus different from digitally accessible information. Structured data is machine-readable and more interoperable. See Table 3 for a shortlist of machine-readable formats.

JSON	Shapefile	RTF (for text)
XML	GeoJSON	HTML
RDF	GML	Excel
CSV	KML	PDF (for text)
TSV	WKT	
ODF	KMZ	

Table 3: Machine-readable formats

The 3-Star Stage: Using non-proprietary formats

Non-proprietary means: not bound to specific software or a specific vendor. Example given: an Excel file (.xls) might seem very open, but it is not. It is bound to Microsoft Excel. This means that everyone that is not in the possession of Microsoft Office is unable to open this file. We call these files being of proprietary formats.

> **Recommendation: Aim to reach 3-star or higher quality data. It is a process. Do not start with 5-stars. Begin with the quick wins. Any star is good to start with.**

It is widely promoted to convert proprietary and non-machine readable files into open and machine-readable formats in order to get a high-quality Linked Dataset.

Machine Readable	Geodata Machine readable	Less readable	Closed
JSON	Shapefile	PDF (For text)	Images (PNG, JPG)
XML	GeoJSON	HTML	Charts
RDF	GML	Excel	
CSV	KML	Word	
TSV	WKT		
ODF			

Table 4: Technical Openness of files

The 4-Star Stage: Use URIs to denote things

If you publish your data as 4-Star Open Data, you will use URIs to denote things and create a first step towards Linked Data. In practice, this means you will convert your files to RDF format and/or you will enrich your metadata with URIs. This is the first step towards Linked Data.

EUROPEAN
DATA PORTAL

The 5-Star Stage: Link your data to other data to provide context

This is a very advanced stage of Open Data. In this stage, the data is linked to other data in order to provide context. This will lead to very interoperable and easy discoverable data. For more information and examples of all stages of the 5-Star Data model, go to **http://5stardata.info/**.

4.1.4. Preparation: legal openness

The data is re-usable in terms of quality; it is technically open. Now it is time for the final preparation step: legally opening up the data. If the data is not legally open, it has no legal right to be re-used. The re-user cannot use the data at all. Legal openness is the basic principle of Open Data.

The implementation of Open Data must be in-line with current legislation, and datasets should be published under an open licence as discussed already in the chapter about licensing. These legal implications are of major importance for any stakeholder trying to make use of Open Data.

Your policy should clearly establish a licensing procedure and take into account national and supra-national legislative matters. Next to the presence of legal aspects in your policy, every dataset should be published individually using a licence.

> **Recommendation: Check the licensing assistant on the European Data Portal for more guidance to choosing a licence.**

Consult the legal department of your organisation to make sure your data is legally open or to check if your policy is compliant. It is the responsibility of the publishing organisation to be up to date with all legislative and legal rules.

For more information about why you need to licence, please also consult the online training module about this topic: **http://europeandataportal.eu/elearning/en/module4/#/id/co-01**

> *Summary of the eLearning Module: Why do we need to license?*
>
> In order for data to be open, it should be accessible (this usually means being published online) and licensed for anyone to access, use and share.
>
> In this module we'll explore the following:
> - Why Open Data needs to be licensed
> - How licences unlock the value of Open Data
> - What type of licence suits Open Data?
> - How to provide for Open Data licensing in the tender, procurement and contracting lifecycle
>
> Even in cases where data has been made available as a public domain dedication without conditions on reuse, an explicit statement is required together with the data to provide users with legal clarity.

EUROPEAN DATA PORTAL

4.1.5. The Final Check

For a final check, use a dataset preparation checklist:

- Check the dataset on quality
- Check the data on timeliness and consistency
- Check the dataset on the use of standards
- Add metadata
- Check if the metadata is described as Linked Data
- Check the dataset on the technical openness
- Check the dataset on legal openness. If it is not open, choose an appropriate licence and apply it to the file.
- Provide licence information and information about the origins

Dataset Preparation Checklist

Provide licence information and information about the origins

Check the dataset on quality

Check the dataset on timeliness and consistency

Check the dataset on legal openness. If it is not open, choose an appropriate licence and apply it to the file

Check the dataset on the use of standards

Check the dataset on technical openness

Add qualitative Metadata

Check if the metadata is described as Linked Data

Figure 33: Dataset Preparation Checklist

4.2. Publishing data

After the preparation phase, the dataset is ready to be published. After this step, the data is available on the web and can be re-used by anyone and for any purpose. As the publishing phase is highly depending on the specific situation of the organisation, we present three short examples of publishing practices: a web publication (single sets and bulk), a portal and an API.
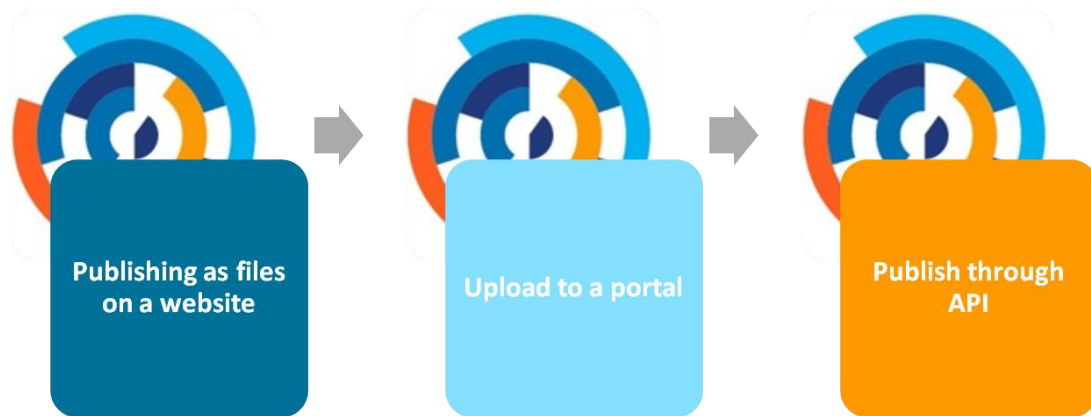
Figure 34: Examples of how to publish data

4.2.1. Publishing as files on a website

Some organisations or portals have a limited number of datasets to share and they do that by publishing the files on their website. For example, the Open Data portal of the German city of Düsseldorf includes a couple of webpages to share data on their city website as shown in the figure below.
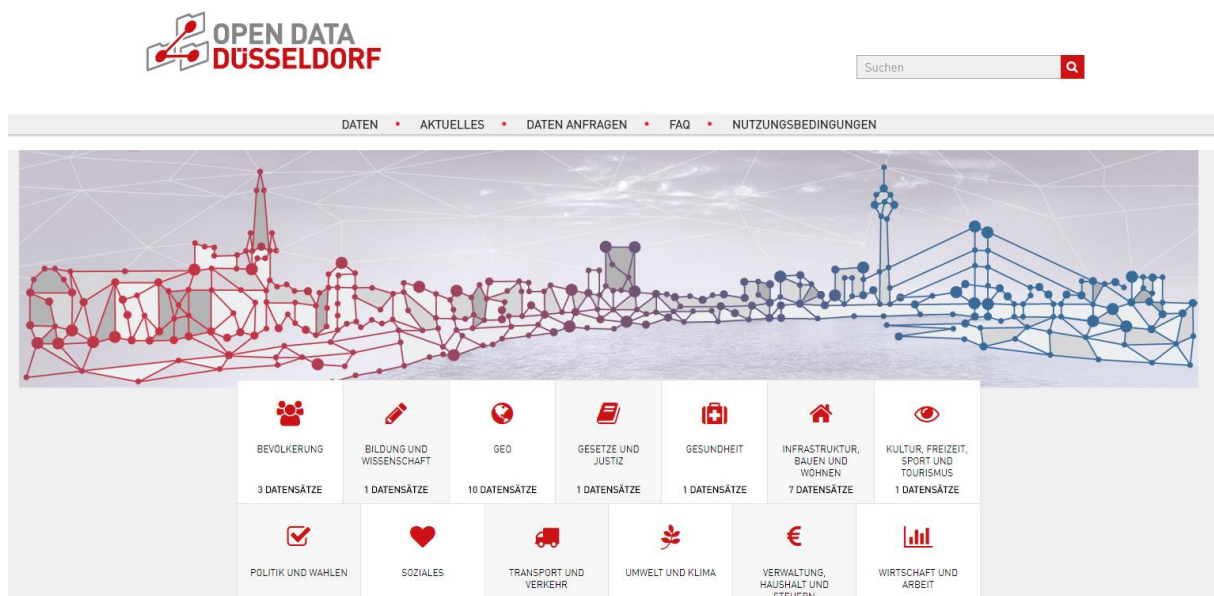
Figure 35: Open Data Portal Düsseldorf (19/12/2017) https://opendata.duesseldorf.de/

4.2.2. Upload to a portal

Upload to a portal is the most used channel for publishing Open Data. An example of how to publish data using CKAN is provided. The figure below shows a screenshot of the CKAN portal on which the data is uploaded. This is the manual CKAN interface, which has some restrictions such as no reusability

of similar templates and no multilingual support. For a full demo, please go to **http://ckan.org**. More information around technical solutions is included in **"Appendix 5 - Technical Solutions".**
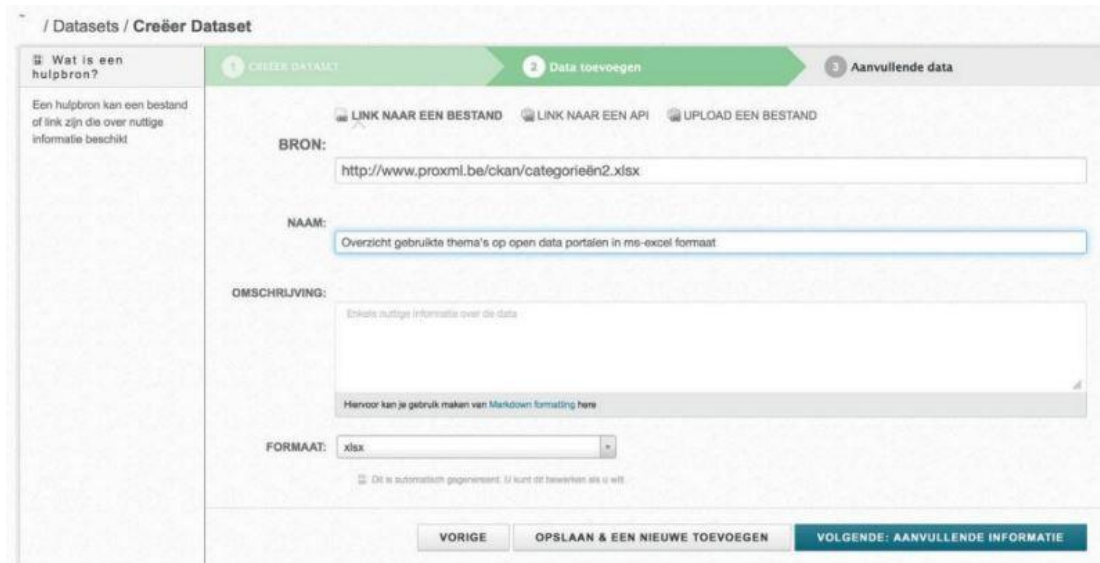
Figure 36: Example of a dataset added to the Flemish data portal using the CKAN interface

4.2.3. Publish through an API

Publication via an API is highly depending on the type of software you are using. However, it is very important to publish the specifications with regard to the API onto your website. A nice example of this case is from the UK, who added an icon to their website to get the API link.
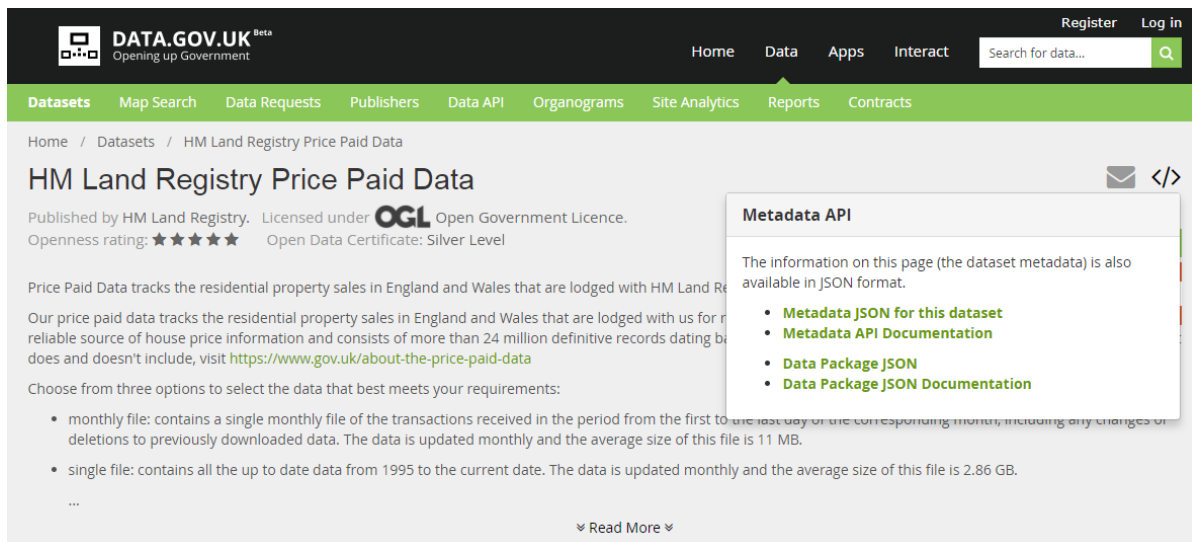
Figure 37: The UK Data Portal provides an API link via this icon on data.gov.uk

4.3. Maintaining data

Data can change over time. Historical data will remain stable, whereas recent data can change frequently. Therefore, we recommend you have a data maintenance process in place. This process consists of maintaining data and metadata regularly, checking URIs & URLs, checking user feedback and continuous improvement and evaluating your success.
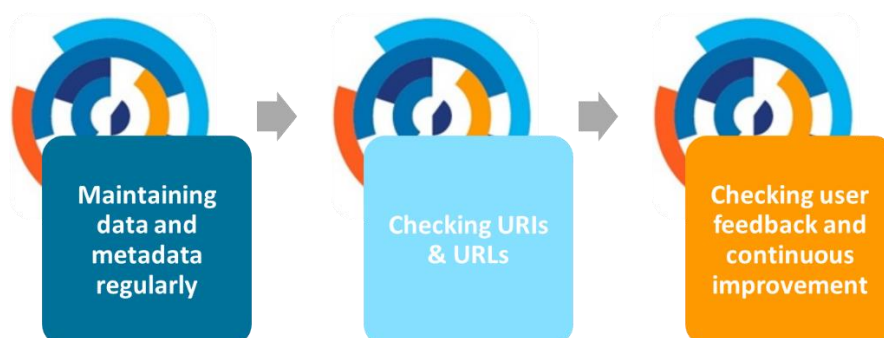
Figure 38: Data maintenance process

4.3.1. Maintaining data and metadata regularly

Both data and metadata can become outdated. Updates, changes or other influences can cause metadata to become obsolete. This will result in low discoverability and lower quality results for re-users in their search for valuable datasets. Therefore, updates should be made regularly. Depending on how the data is generated, the relevance of updates and the re-use of the dataset (frequency, quality), the dataset needs to be updated with frequent time intervals. Perform a thorough assessment of the data and discuss the policy within your organisation.

> **When an update is made, a key recommendation is to mention the date of the update in the metadata!**

4.3.2. Checking URIs & URLs

The World Wide Web is dynamic. Therefore, it is important to regularly check if all your URIs and URLs (from and to datasets) are still working. If the dataset's URI or URL changes, a website that refers to your dataset will redirect users to non-existing pages.

4.3.3. Checking user feedback and continuous improvement

User feedback will increase the quality of your data publications. Users can provide feedback about the data on any aspect and by incorporating feedback into your processes, the usability and discovery of the data can be improved. Consider adding an option to receive feedback as a potential improvement of the portal.

5. Ensuring and monitoring success

To ensure and monitor the success of your Open Data initiative, it is important to engage re-users, to monitor various key aspects of your initiative. This will enable you to constantly improve your work by acting on the feedback of re-users and learning from your key monitoring indicators.
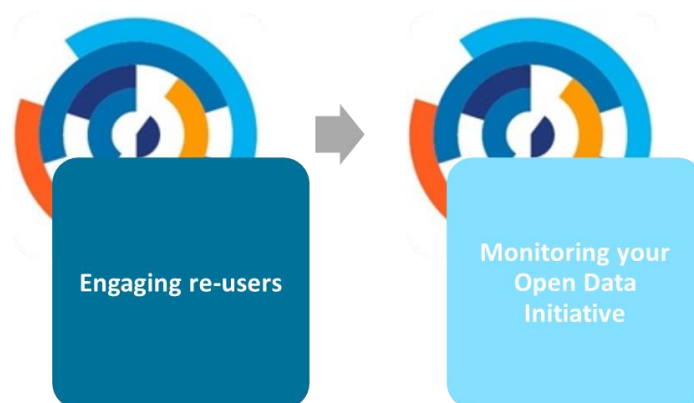
Figure 39: Ensuring and Monitoring Success

After publishing the data and having your lifecycle in place, it is time for the last step: Evaluating the success of your implementation. Your experience is a great source of improvement. After thoroughly evaluating your efforts, metrics and the benefits, revise your policy and your strategy and adapt where necessary. From what you have learned, what can be improved? Formulate next steps and implement them. From there on you can start the Open Data Lifecycle and keep the work in motion.

A first step to measure your success is to engage re-users. Your stakeholders will play a key role in underlining the benefits and concerns of your Open Data activities. A second step relates the monitoring of your Open Data initiative.

5.1. Engaging re-users

Publishing Open Data is not only about making the data accessible on the web. You can make your Open Data initiative a larger success if you engage the re-users. Going beyond the approach of simply publishing data online, Tim Davies has developed a five-star Open Data engagement model. The model developed explores how to

- Be demand driven
- Put data in context
- Support conversation around data
- Build capacity, skills and networks
- Collaboration on data as a common resource
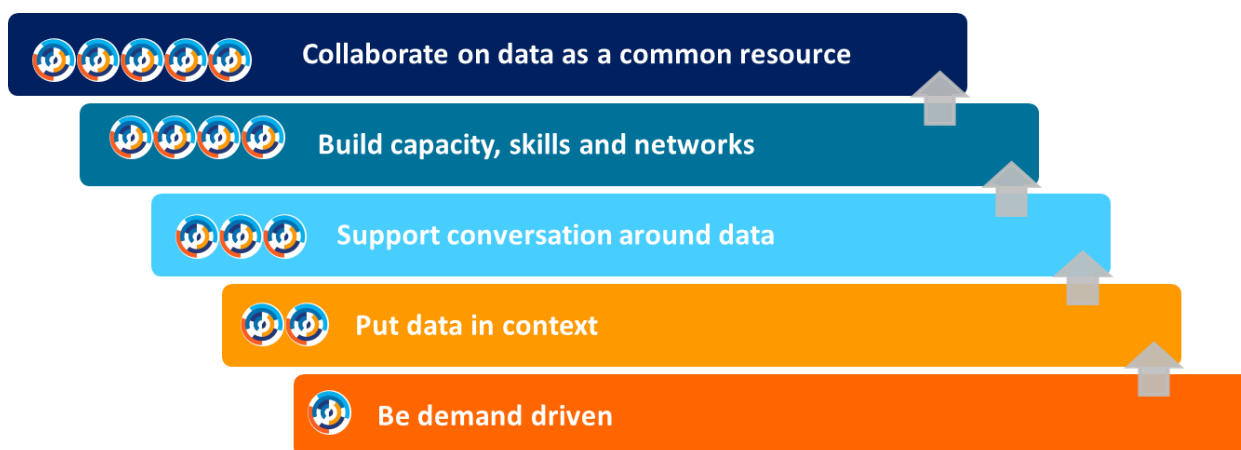
Figure 40: Tim Davies' five star Open Data engagement

To learn more about the Open Data engagement model go to **"Appendix 4 - Open Data engagement model".** To learn more about the skills needed to work with Open Data, read the EDP report on e-skills and Open Data: *https://www.europeandataportal.eu/sites/default/files/edp_analytical_report_n2_-_e-skills.pdf*

5.2.   Monitoring your Open Data initiative

To monitor the success of your Open Data initiative, consider implementing metrics to your publications in order to evaluate its success. With these metrics, you can evaluate several indicators. The most useful evaluation activities are performance of the data, performance of the system, and collection and preparation performance.
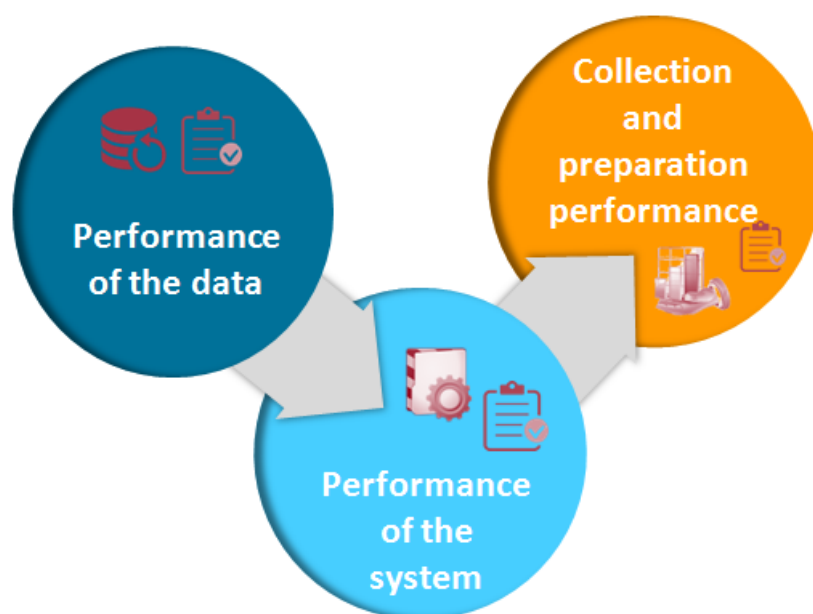
Figure 41: Evaluation activities

Performance of the data

This evaluation includes checking the number of downloads and page views. These are not the same, but both indicate the popularity of the dataset. It does not indicate the usefulness of the dataset: one cannot conclude whether the data has been re-used based on the number of downloads.

Performance of the system

An important metric, especially when the data is available through an API. Here you want to evaluate whether the system can handle the requests, if there has been any downtime, and if there are performance consequences for other systems.

Collection and preparation performance

To evaluate user feedback, the usefulness of datasets is used. Usefulness is an indicator caused by the qualitative usefulness (is it helpful for a particular purpose?) and the practical usefulness (is the data described, clean, dense enough, etc.). The latter is an indicator you can influence, as this reflects the performance of the Open Data Lifecycle.

You should consider including metrics that will enable you to measure the success of the publication of data and your metadata. Think of the following metrics:

- Qualitative Feedback
- Number of downloads per set
- Click through rate
- Re-user rating of quality
- Cost per download

Free tools such as PIWIK are useful for these analyses. Some data portal software solutions come with built-in metrics.

This is an example of the French Open Data portal where re-users give feedback on the datasets. It shows how many times it has been re-used and how many followers there are.

References

Berners-Lee, T. (2015) Linked Data; http://www.w3.org/DesignIssues/LinkedData

Berners-Lee, T. (2013) 5 Star Open Data; http://5stardata.info/

Broad, E; Smith, F; Duhaney, D; Carolan, L (2015) Open Data in Government: how to bring about change; http://theodi.org/open-data-in-government-how-to-bring-about-change

Cabinet Office (2013) G8 Open Data Charter and Technical Annex; https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex

Davies, T. (2012) The Five Stars of Open Data Engagement; http://www.opendataimpacts.net/engagement/

EUCLID (2014) Chapter 1: introduction and application scenarios; http://www.euclid-project.eu/modules/chapter1

EUR-Lex (2013) Directive 2013/37/EU; http://eur-lex.europa.eu/legal-content/NL/TXT/?uri=CELEX:32013L0037

European Union (2015) EU Open data - The basics for EU data providers

Feigenbaum, L. (2015) SPARQL by example; http://www.w3.org/2009/Talks/0615-qbe/

Herreweghe, N. van (2015) Open Data Manual; https://www.w3.org/2013/share-psi/wiki/images/b/bb/Open_Data_Handbook_12022015_EN.pdf

Keyzer, M. de, Loutas, N., Goedertier, S. (2013) Introduction to RDF & SPARQL, slide 9; https://joinup.ec.europa.eu/community/ods/document/tm13-introduction-rdf-sparql-en

Kolodziejski, M. (2013) Digital Agenda and the Economic Development of European Regions; http://www.europarl.europa.eu/RegData/etudes/divers/join/2013/513984/IPOL-REGI_DV(2013)513984_EN.pdf

Lapsi-Project (2013) The PSI Directive vs Generally Acknowledged Open Data Features; https://ec.europa.eu/digital-single-market/en/news/legal-aspects-public-sector-information-lapsi-thematic-network-outputs

Open Knowledge (2015) Open Definition V2.0.; http://opendefinition.org/od/

Rogers, K (2015) Improving government access to government data.; http://opendatahandbook.org/value-stories/en/improving-gov-access/

W3C Foundation (2015) Data on the Web Best Practices; http://w3c.github.io/dwbp/bp.html#metadata

Appendix 1 - The PSI Directive vs. Generally Acknowledged Open Data Features

(Lapsi-project, 2013)

Confronting the PSI Directive provisions and the widely acknowledged Open Data features would lead to point out that:

1. PSI refers to "documents held by public sector bodies". While the PSI Directive encourages public sector bodies to make any of their documents - and data - available for re-use, it also sets some access and re-use restrictions to such documents. First, the Directive does not contain an obligation to allow re-use, thus leaving each EU Member State or public sector body to decide themselves whether a document should be reusable or not. Second, the Directive does not change the national rules for access to documents, so that each EU Member State could maintain its own access restrictions (usually due to privacy or national security concerns). In addition, the PSI Directive currently does not apply to documents held by public service broadcasters, educational and research establishments, and cultural establishments. Open Data refers to "data" as a potentially much broader term which may involve any kind of work, knowledge, data or information with no given source limitations. Access restrictions are conceived mainly for data affecting privacy, confidentiality or public security.

2. PSI can be made available charging a price for re-use. The PSI Directive sets the charging upper limit at the recovery of total costs of collecting, producing, reproducing and disseminating documents together with a reasonable return on investment, though leaving the right to ask for lower charges or no charges at all. In addition, the Directive encourages making documents available at charges that do not exceed the marginal costs for reproducing and disseminating the documents. Open Data should be available at no more than a reasonable reproduction cost. Yet, the online availability without charge is the first-choice option.

3. PSI itself does not affect the existence or ownership of intellectual property rights of public sector bodies: while public sector bodies might be encouraged by the Directive to exercise their copyright in a way that facilitates re-use, the default rule adopted by the Directive seems to be the traditional all rights reserved copyright rule. Therefore, should a public-sector body have any intellectual property right on its information, it is up to the public sector body itself to decide how broadly its information has to be licensed. Open Data experts specifically require data to adopt an Open Licence (e.g. Creative Commons, Open Government Licence) in order to be disseminated in a truly open fashion, thus aspiring to some rights reserved copyright rule.

Appendix 2 - Master Data Management Change Plan

(Herreweghe, N. van, 2015)

1. **Identify Open Data master data**. Based on several criteria, data identified as not being master data is identified and therefore not included in an Open Data management process.
2. **Identify source systems**. What is the origin of the master data and its metadata? Which source systems do they produce?
3. **Collect and analyse metadata on master data**. Refer to the chapter about metadata for Open Data, which describes the necessary fields.
4. **Appoint data stewards**. These individuals have expertise in both current source systems and Open Data to make the same rules apply to all data sources.
5. **Draw a data governance programme and establish a data governance council**. The programme defines how, where, and with which definitions master data is established. The data governance council decides in consultation which normalisation procedure is used.
6. **Develop a master data model or logical data model**. Depending on the available databases and data warehouse (if applicable) and the required distribution of the information, a logical and physical data model is designed to be managed under the MDM process.
7. **Consider a tool**. If high volumes of data are managed, we recommend using an MDM toolset.
8. **Design a supporting infrastructure**. For bodies managing large volumes of data and aiming to open up data automatically, consider using a supporting infrastructure for implementing Extract, Transform, Load (ETL) processes.
9. **Generate and test master data**. Check the master data quality and consistency during manual or automated inspections. It is impossible to make and keep all master data accurate in one go. ETL toolsets often contain possibilities for providing this. However, in some cases specific tests may be needed (for instance to anonymise Open Data).
10. **Implement maintenance processes**. Processes are never static, and the management of MDM and Open Data streams is not either. Therefore, you should provide a process for maintaining metadata and ETL functionality to maintain the quality of the data.

Appendix 3 - The Extract Transform Publish (ETP) Process

(Herreweghe, N. van, 2015)

Publishing data as Open Data overlaps with the existing data warehousing process called ETL (Extract, Transform, Load). It is convenient to leverage this process as a blueprint, as the techniques are in place. Thus, no new technique needed.
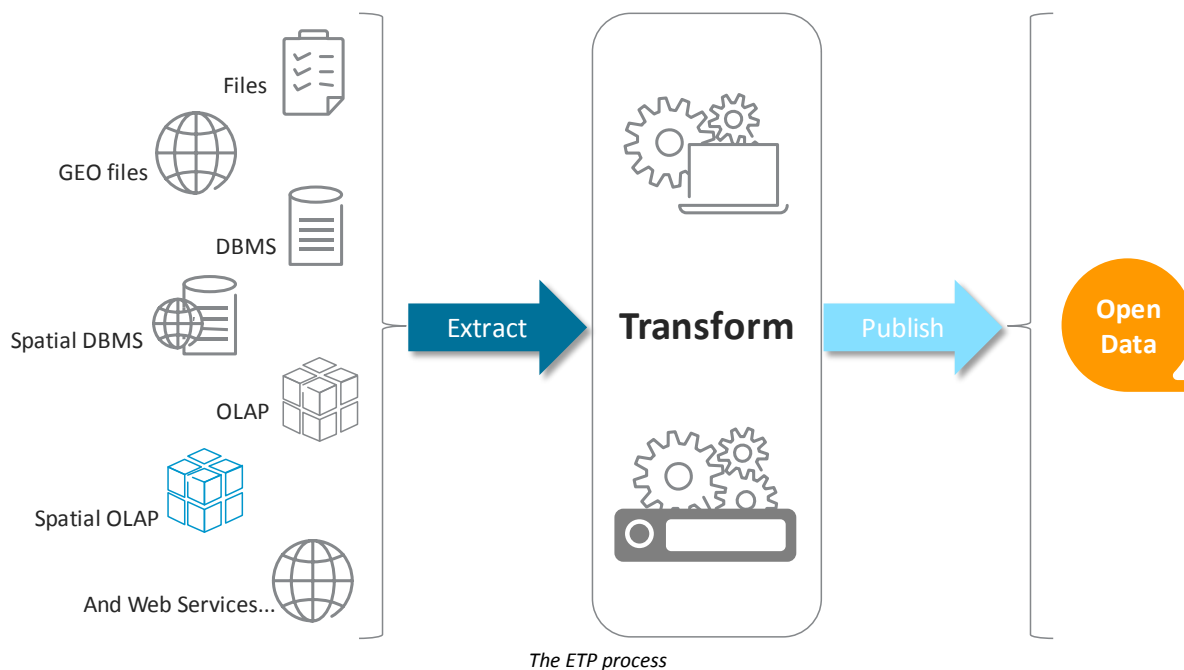
The ETP process

The ETP-Process has three steps:

1. Extract

Data can be extracted from all kinds of sources that include newly generated data or data from another internal or external source. Depending on the data management structure of the organisation, this step in the ETP process differs.

2. Transform

In this process, data should be transformed into Open and Linked Data. This 'Preparing' or 'Cleaning' data phase should be a standard process within the organisation. A policy should be in place that sets the standard for the organisation with regard to their guidelines for preparing data before publication.

3. Publish

It is possible to publish data through various channels:
- Via download (single or dump file) on (existing) website
- Via a portal
- Via an API

Publishing Data from different starting points

The starting point of publishing data depends on the data management structure of the organisation, but will always have the characteristics of the 'Extract' step of the ETP-process. In order to give an indication of different possibilities, various scenarios are provided as examples for these starting points. These are:

a) From an existing publication
b) From an existing dataset
c) From a database
d) From an existing source system or package
e) From different sources – consolidating data

a) Starting from an existing publication

In this scenario, we start from an existing process during which data is collected and processed before it is included in a publication. Think of all the publications made available by public sector and the amount of data generated. Data is the result of a number of process steps that are carried out by one or more administrations.

Because the government will continue to issue publications, and administrations will continue to collect data for publication, this seems to us the easiest scenario for Open Data. The published data can be transformed into an Open Data stream using those scenarios:

- For new publications; in this case, an additional step will be included in the process to make the data also available as Open Data
- For existing publications; in this case, the existing process will be adapted to create the Open Data stream and to keep it up to date

Once you have set up the process for extracting an Open Data stream from the publication and publishing it, you can apply this repeatedly. This way, the Open Data stream is kept synchronized with every update of the publication. The essence is that an additional step is included in the process to publish the data also as Open Data. See figure below: In the existing process, the IT department will have to identify the moment when the data are ready to be taken to the Open Data publication step. In this case, we only have a number of 'publish' activities to perform.

1. Extract
N/A.

2. Transform
N/A.

ETP-Process with existing publication
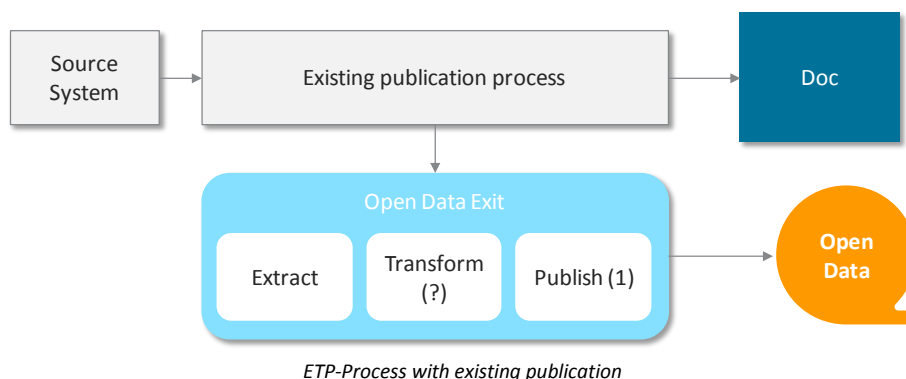
3. Publish

Once the data is ready for publication as Open Data, the following steps remain to be completed:

- Collecting metadata
- Publishing dataset (preferably automatically)
- Choosing licence model
- Offering necessary conversions on the platform, and possibly also an API
- Setting up a feedback loop, making sure contact details of the public-sector organisation are available in case of comments
- Ensuring regular updates

b) Starting from an existing dataset

Many organisations currently publish a fair amount of information on websites in an open and downloadable format like CSV (note: if the data is in closed formats (e.g. PDF, XLS), see scenario 1). We presume that the organisation will keep making information available via downloads or viewers, and that Open Data is an additional channel for releasing this information. This means that the underlying process will continue to exist and may be extended by an additional step. Usually, public data does not entirely meet the Open Data criteria. With minimal efforts and additional steps, this could prove a sound basis for quickly turning them into Open Data.

This scenario considers a few additional steps, since the existing process deals with the data in a very different manner. These steps are:

1. Extract

Isolate data and filter them from the database in a uniform dataset. This may require an additional step to extract data directly from the database. We may also select different data for the dataset (e.g. fewer fields, anonymised) than in the existing process. Of course, this step will not be necessary if the data from the publication already meets the Open Data criteria.

2. Transform

Before publishing data as Open Data, transform the data to Open and Linked Data. Therefore, it is necessary to go through the 'preparation' step of the data.

3. Publish

The following steps are always required in this scenario, once the dataset has been created.

- Collecting metadata
- Publishing dataset (preferably automatically)
- Choosing licence model
- Offering necessary conversions on the platform, and possibly also an API
- Setting up a feedback loop, making sure contact details of the public sector organisation are available in case of comments
- Ensuring regular updates

c) Starting from a database

In many cases, the core data is included in a database created for an application to support a business process for the public sector organisation. This is also the starting point of this scenario: to extract data

from a database and transform it into an Open Data stream.

Most databases use a structure for storing data that is less suitable for Open Data. The main aim is to transfer data to the application to be able to create, retrieve or delete data. This structure is called OLTP (= online transaction processing) and relational in nature: data is linked through a relation (key) in different tables. In summary, the basis of this scenario is that the data is extracted from the application database first, before it is prepared for publication as Open Data. An additional processing step is needed to extract the data and transform it or make it consistent for Open Data publication. In this case, transformation tools are necessary as well.
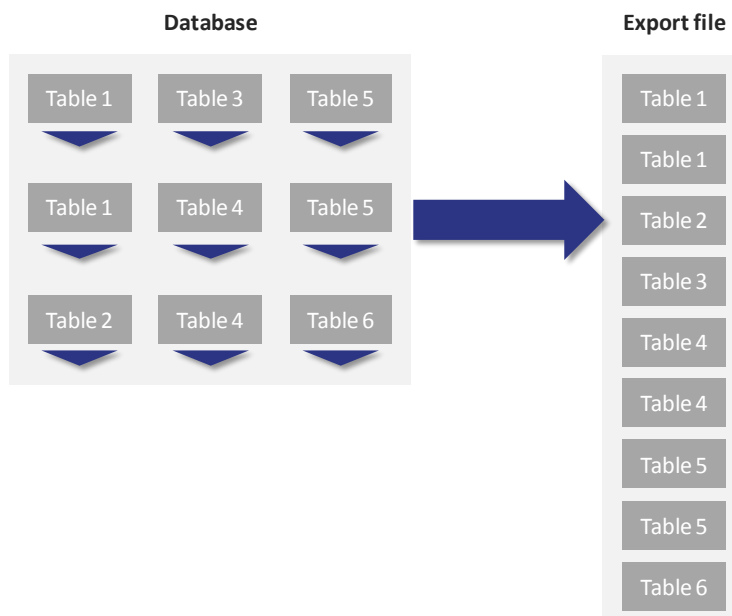
Extracting from a Database

The assumption in this scenario is that we have an application which is developed internally on one of the existing internal environments (for instance, built in Java, .NET or APEX) and the database is one of the standards used by the public sector organisation (like Oracle, SQL server). If this is not the case, we propose proceeding with the next scenario. We also assume that the previous scenarios are not applicable as starting point. In other words, data will first have to be extracted from databases before it is opened up to the Open Data Platform. In this scenario, the techniques 'extract' and 'transform' are very useful, because there is no existing process to start from or hook into. Therefore, we start directly from a database (through a query or tools).

A remark to be made in this context is that we do not equate this type of 'technical' operations with the 'content related' operations which usually take place during the 'transform' step. We refer to complex transformations (e.g. linking consecutive enrolments of pupils from different systems with each other (i.e. lower, secondary, higher, VDAB, etc.) to construct an entire study career). This may require more than one technical step. We propose the following steps:

Mapping
First, create a logical data map in which the physical relations of the database are ignored as much as possible. The result is that an extract from the database will almost never be published as Open Data directly. However, it is possible. In most cases, it will be impossible to publish a one-to-one copy and

steps will have to be taken to transform the database copy into a stand-alone format (for instance replacing keys by values, entering references, making references consistent, M = Man, etc.). Furthermore, steps are needed to combine tables from the database with each other until you arrive at the logical data map. Transformation technology may possibly also have to be used to make the data consistent.
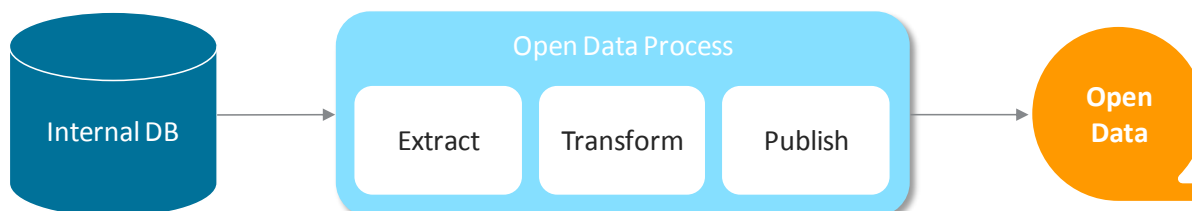
The ETP-Process starting from an existing database

Extract

Most database systems have standard techniques for reading out tables and making them available as flat files. For example:

- Oracle: via EXPORT tool
- Microsoft SQL Server: Import and Export Wizard MySQL: via mysqldump tool
- PostgreSQL: SQL Dump procedure

When data changes frequently, a good alternative may be to write a programme that reads out the data from the Database Management System (DBMS) via ODBC of JDBC drivers.

ODBC (or direct SQL, which is in fact the same thing) will allow programming of more complex extraction logic. Finally, the possibilities provided by each ETL tool set can be used. From a practical point of view, it is also recommended for complex transformations to store the data, either temporarily or permanent, in a database which can then be further used to realize the final Open Data stream. Two dimensions are always important in such procedures:

- Exporting the entire content of the database and publishing it as Open Data
- Downloading the modifications /delta with respect to the previous version and combining it with the Open Data stream

Transform

This encompasses a thorough quality check of the data, as is the case in every data warehouse environment. For instance, using uniform names for fields and content instead of cryptic abbreviations, not using 0 or 1 for gender but M = Man, storing addresses in a consistent manner, writing names in full and in the same format, etc. Because we cannot assume here that a process is already in place, all these transformation steps need to be carried out. At this stage, content can also be processed like anonymising data or combining datasets to achieve a uniform granularity. It is preferred to export the extracted data as quickly and frequently as possible to Open Data to allow users and citizens to have the latest reference data at their disposal, especially when it concerns rapidly changing data. Consequently, it is important that the transformations that these data undergo are reproducible and should preferably happen automatically.

Publish

Once the data are ready for publication as Open Data, the following steps remain to be completed:

- Collecting metadata
- Publishing dataset (preferably automatically)
- Choosing licence model
- Offering necessary conversions on the platform, and possibly also an API
- Setting up a feedback loop, making sure contact details of the public sector organisation are available in case of comments
- Ensuring regular updates

d) Starting from an existing source system

The starting point in this scenario is that the public sector has one or more operational systems or packages from which Open Data can be obtained, provided a number of operations are performed, either with or without intervention from the supplier or service provider that manages the package for the public sector organisation concerned. In this scenario, the additional complexity is that a package is involved that may not allow direct access to the database.

This scenario applies when the data is included in a (commercial) package, in which it is sometimes impossible to access the database directly. This scenario also applies when the public sector organisation involves a service from a third party (like software-as-a-service) and the service is provided externally (like a Cloud application).

This scenario is an extension of the previous scenario, but differs from it in that the database cannot be accessed directly, as shown in the figure below. The advice given in case of a package is to never copy and publish data one-to-one. Databases and field names are often described cryptically and the database structure is designed in such a way that it is only optimized for online transactions. In case of upgrades the structure usually changes, which means you have to start all over again.
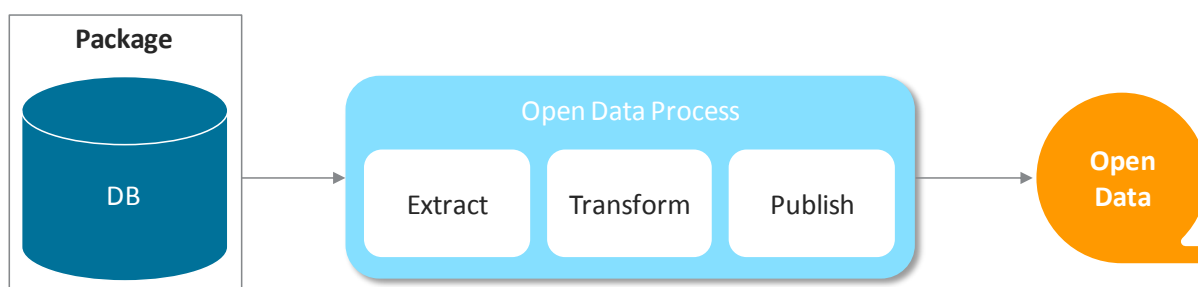
The ETP-Process when starting from an existing source system

The following steps are relevant as well:

Extract

We assume that the public sector organisation can discuss with the supplier of the package how the data can be opened up:

- Through a procedure belonging to the package. Package suppliers often deliver a programme or script for reading out the data, even including specific parameters. Since this procedure is written and maintained by the supplier (for instance, in case of

package upgrade), you are certain that this procedure is forward and backward compatible.

- Through APIs, if provided by the package. This means you will have to write a programme that uses the APIs to obtain data. Please note that some APIs can also carry out operations on the data before delivering them, like consolidation, aggregation, etc.
- Through a new programme or script which directly reads out the data from the package database. However, in this case one is dependent on the database design of the supplier, which tends to change with each new version or upgrade. This means that you will have to adjust the programme each time. Therefore, this approach is not recommended.

Two dimensions are always important in such procedures:

- Exporting the entire content of the database and publishing it as Open Data
- Downloading the modifications /delta with respect to the previous version and combining it with the Open Data stream

Transform

This encompasses a thorough quality check of the data, as is the case in every data warehouse environment. For instance, using uniform names for fields and content instead of cryptic abbreviations, not using 0 or 1 for gender but M = Man, storing addresses in an consistent manner, writing names in full and in the same format, etc. Because we cannot assume here that a process is already in place, all these transformation steps have to be carried out. At this stage, content can also be processed like anonymising data or combining datasets in order to achieve a uniform granularity.

Publish

Once the data are ready for publication as Open Data, the following steps remain to be completed:

- Collecting metadata
- Publishing dataset (preferably automatically)
- Choosing licence model
- Offering necessary conversions on the platform, and possibly also an API
- Setting up a feedback loop, making sure contact details of the public sector organisation are available in case of comments
- Ensuring regular updates

e) Starting from different sources – consolidating data

This is a scenario for experienced bodies involving Business Intelligence and data warehouse techniques. It expects the public sector organisation to already have some data warehouses and reporting environments and be greatly experienced in opening up data from different sources. All 'Extract' and 'Transform' techniques and tools are already in use and can be re-used here. Data published in a data warehouse is often the result of a series of operations, in terms of both content and aggregation. First, the relevant data is extracted from different sources and stored consistently in an Operational Data Store (ODS) environment. For Open Data, this has the advantage that this data no longer needs to be extracted separately from the source systems. An ODS is also the lowest form of granularity of the data, and therefore an ideal source for opening up all the data here. After that, data is often processed in phases and stored in staging tables in the meantime. Several programs then run

on this to further aggregate the data and make them consistent before uploading them to a data warehouse. For Open Data we can also start from a staging table from which the Open Data stream can be produced. The great advantage is that in this case all the corrections and operations of the data have already taken place.

(ODS= Operational Data Stage)
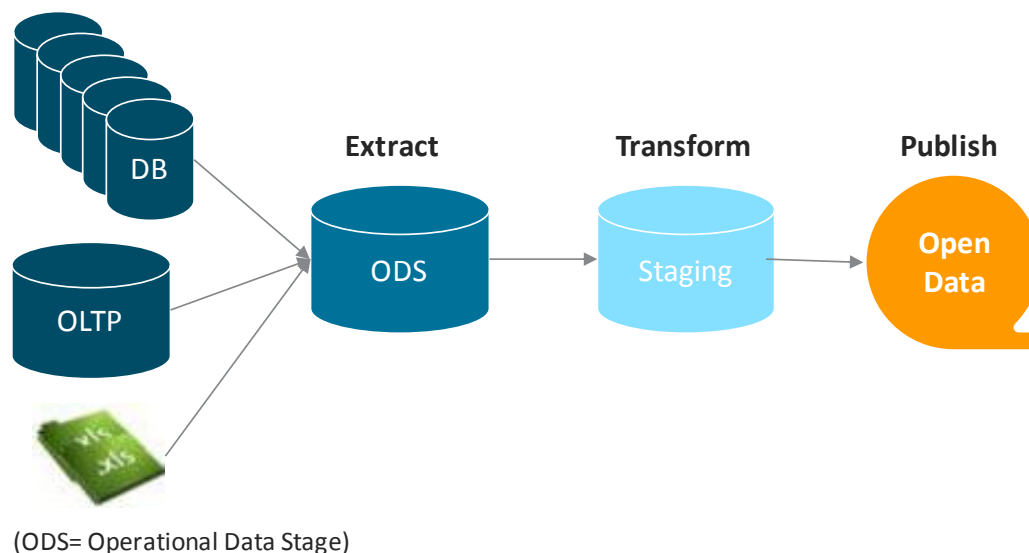
The ETP-Process when consolidating data

In this case, we assume that the Open Data team will make maximum re-use of a number of existing facilities within the public sector organisation.

This scenario is thus valid when an ODS environment is already in place and/ or a number of staging tables are available as starting point. The Open Data team can then use the 'Extract' and 'Transform' programs to make consistent Open Data streams immediately. However, the Open Data team must in this case also check or adjust the aggregation and granularity in order to keep the Open Data streams as fine-grained as possible. A separate or specific staging table for Open Date may therefore have to be made. In this case, an extra programme should be provided within the public sector organisation's technology. When complex transformations are carried out on the source data, staging tables are added to combine data from different bodies and, in the end, publish one consolidated file (for instance one address file for all bodies with a similar data model). The following steps are relevant: We propose to re-use existing standard ETL environments as much as possible and technologies for Open Data. We do not see the need to introduce other tools.

Extract
In this case, we assume that all extract functionalities are available within the existing BI or DWH and we do not specifically have to set it up for an Open Data stream.

Transform
Again, we start from the existing BI or DWH environments. All Open Data streams can be taken from the latest staging tables. A specific staging table may just have to be drawn up for Open Data, but this should be examined with respect to the granularity.

Publish

Once the data are ready for publication as Open Data, the following steps remain to be completed:

- Collecting metadata
- Publishing dataset (preferably automatically)
- Choosing licence model
- Offering necessary conversions on the platform, and possibly also an API
- Setting up a feedback loop, making sure contact details of the public sector organisation are available in case of comments
- Ensuring regular updates

EUROPEAN
DATA PORTAL

Appendix 4 - Open Data engagement model

(Davies, T. 2012)

Rating on the engagement scale	Description
★ - ONE STAR **Be demand driven**	⊙ Are your choices regarding the kind of data you release, how it is structured and the tools and support provided around it based on community needs and demands? ⊙ Have you got ways of listening to people's requests for data, and responding with Open Data?
★★ - TWO STARS **Put data in context**	⊙ Do you provide clear information to describe that data you provide, including information about frequency of updates, data formats and data quality? ⊙ Do you include qualitative information alongside datasets such as details of how the data was created, or manuals for working with the data? ⊙ Do you link from data catalogue pages to analysis of the data that your organisation, or third-parties, has already carried out with it, or to third-party tools for working with the data?
★★★ - THREE STARS **Support conversation around data**	⊙ Can people comment on datasets, or create a structured conversation around data to network with other data users? ⊙ Do you join the conversations? ⊙ Are there easy ways to contact the individual 'data owner' in your organisation to ask them questions about the data, or to get them to join the conversation? ⊙ Are there offline opportunities to have conversations that involve your data?
★★★★ - FOUR STARS **Build capacity, skills and networks**	⊙ Do you provide or link to tools for people to work with your datasets? ⊙ Do you provide or link to 'How To' guidance on using Open Data analysis tools, so people can build their capacity and skills to interpret and use data in the ways they want to? ⊙ Do you go out into the community to run skill-building sessions on using data in particular ways, or using particular datasets? ⊙ Do you sponsor or engage with capacity building to help the community work with Open Data?
★★★★★ - FIVE STARS **Collaborate on data as a common resource**	⊙ Do you have feedback loops so people can help you improve your datasets? ⊙ Do you collaborate with the community to create new data resources (e.g. derived datasets)? ⊙ Do you broker or provide support to people to build and sustain useful tools and services that work with your data? ⊙ Do you work with other organisations to connect your data sources?

Tim Davies' 5-star Open Data Engagement Model (Davies, T. 2012)

EUROPEAN
DATA PORTAL

Appendix 5 - Technical Solutions

Below is an overview of technical solutions to use for the implementation of your Open Data initiative. Furthermore, there are a number of re-usable implementations or components of implementations of Open Data portals that can be re-used free of cost, such as the European Open Data Portal itself.

CKAN

Developed by Open Knowledge, CKAN has become one of the major standards within Open Data portals. Many of the known and governmental Open Data portals, such as the European's own data portal, are CKAN based portals. As CKAN is open source, it is continually improved and is available free of charge. The catalogue system comes with many strong features such as harvesting, publishing and auditing and has integrated data storage. It interoperates with many of the technical standards supported by the EU. When you use CKAN, you need a CMS functionality, which means that a separate content management system is needed. Recommended are Drupal, Wordpress, and Django. Go to **http://ckan.org** for more information and a live demo.

DKAN

Developed by Drupal, DKAN is yet another open source Open Data Catalogue system, which is recommended frequently and meets the US project Open Data requirements. This software is a highly complementary offering to CKAN. However, DKAN is Drupal based and comes with an integrated content management system, which can be integrated easily into other content management systems. For a live demo of DKAN, go to **http://demo.getdkan.com/**

Openrefine

A dataset analysis and cleaning tool. It is a Google tool.
http://openrefine.org/

ODI Certificate Tool (beta)

The Open Data Institute has developed a self-assessment tool that tests the maturity of datasets on various aspects with regard to licences and openness. Evaluate which steps you still have to take to allow others to find your datasets more easily.

https://certificates.theodi.org/
The Datatank

The DataTank is open source software, just like CKAN, Drupal or Elastic Search, which you can use to transform any dataset into an HTTP API.

http://thedatatank.com/
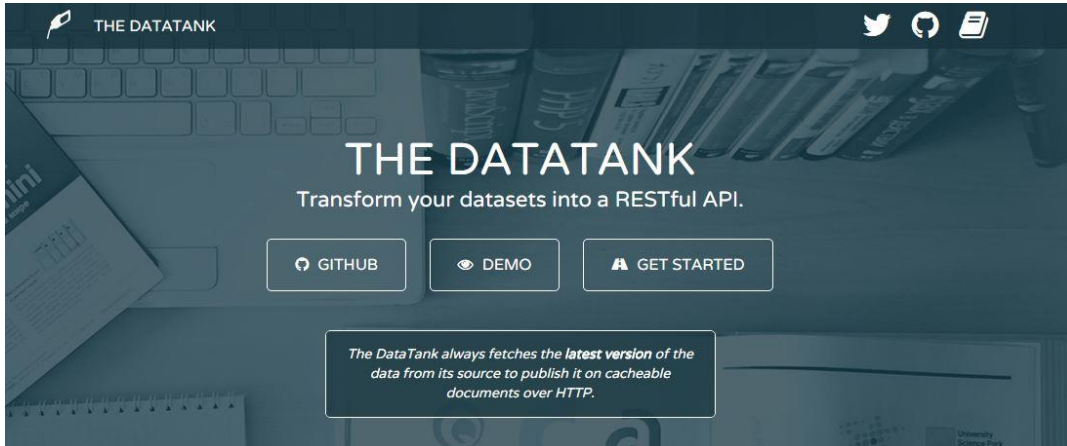
Snorql

This tool is useful for learning SPARQL Queries

http://snorql.nextprot.org/
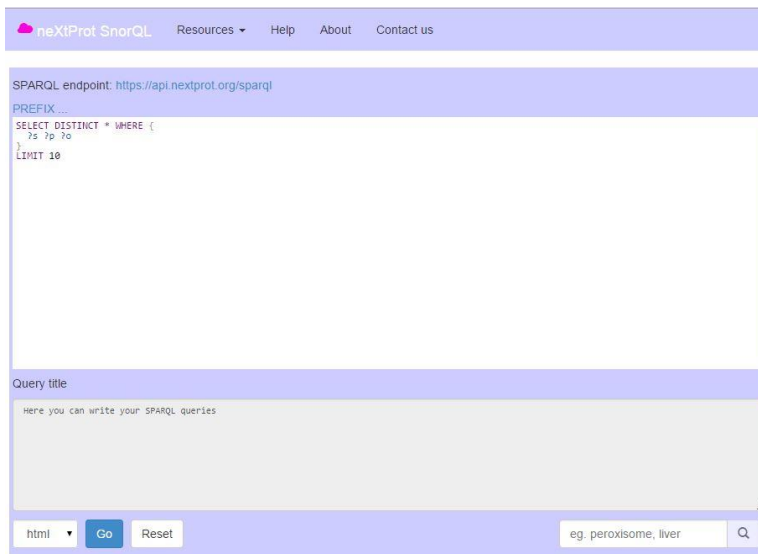
Appendix 6 - Online training material

In addition to the online training modules provided on the European Data Portal that are referenced throughout this Goldbook, there are a number of other relevant training resources per topic online that are worthwhile consulting.

Discover the European Data Portal's eLearning

What is eLearning?

Our experts have selected 13 short modules designed for anyone to discover more about Open Data. The modules suit all levels from beginners to experts.

Lesson 1 - What is Open Data?

Open data is data that anyone can access, use and share. Governments, businesses and individuals can use open data to bring about social, economic and environmental benefits.

In this module, we will explore the following:

- ✓ What is Open Data?
- ✓ What is data?
- ✓ What makes data open?
- ✓ Why do we need Open Data?

Lesson 2 - Unlocking value from Open Data

Open data has the potential to help grow economies, transform societies and protect the environment. In this module, we explore how governments, businesses and individuals are using open data to create new value.

**In this module, we will explore the following:**

- ✓ Innovation and growth in businesses
- ✓ Opportunities for governments
- ✓ Impact on society and public policy
- ✓ Benefits for culture and the environment

Lesson 3 - Open data. Agent of change.

The most successful open data initiatives share similar characteristics. Understanding these approaches can help those wishing to unlock the value of open data for themselves.

In this module, we will explore the following:

- ✓ Open data as an agent of change
- ✓ Leadership and engagement
- ✓ Supply and demand
- ✓ Culture change for new markets

Lesson 4 - Why do we need to license?

For data to be open, it should be accessible (this usually means being published online) and licensed for anyone to access, use and share.

In this module, we will explore the following:

- ✓ Why open data needs to be licensed
- ✓ How licences unlock the value of open data

✓ What type of licence suits open data

✓ How to provide for open data licensing in the tender, procurement and contracting lifecycle

Lesson 5 - What makes quality Open Data?

Assessing how usable open data is is not something that can be done quickly. There are a number of community-based standards and quality markers that can help you assess the usability of data.

In this module, we will explore the following:

✓ What makes data usable?

✓ How standards help increase the usability of data

✓ Marques of quality

Lesson 6 - Measuring success for Open Data

Successful open data initiatives do more than simply put data on the web. The most data-savvy organisations also put in place frameworks and policies to support and incentivise innovation. Open data communities need to be built and success stories communicated. Together, these will help more people understand the benefits of open data.

In this module, we will explore the following:

✓ Measuring success

✓ Being demand focused

✓ Keeping on track

Lesson 7 - Why should we worry about sustainability?

Open data must be relevant, up-to-date and accessible in order to be useful. Together, these qualities help make a dataset sustainable. A sustainable programme is one that continues to regularly release data with at least the same or improving quality and quantity. For something to be sustainable it must be able to be maintained at a certain rate or level. For an open data release to be sustainable, it must maintain regular updates with at least the same level of quality and quantity.

In this module, we will explore the following:

✓ How to make sense of sustainability

✓ Why sustainability matters to you

✓ What to look for in sustainable open datasets

Lesson 8 - Getting to grips with platforms

A platform is a major piece of software on which smaller pieces of software and content can be run. For open data, the largest platform is the web. However, lots of other purpose-built software helps simplify publishing open data and provides interactive tools for users to explore.

In this module, we will explore the following:

✓ Recognising open data platforms

✓ Understanding their importance to users

✓ Evaluating the key options

✓ Using platforms to explore data

Lesson 9 - Choosing the right format for open data

The format of an open dataset is the way the data is structured and made available for humans and machines. Choosing the right format enables simpler management and reuse of the data. To maximise reuse of the data, it may be necessary for a publisher to use a number of formats and structures available across different platforms that suit a user's needs.

In this module, we will explore the following:

- ✓ Why formats matter to open data
- ✓ Choosing the correct structure
- ✓ Access open data formats
- ✓ Keeping it simple with CSV

Lesson 10 - How useful is my data?

Assessing how useful open data is can vary depending on the domain and the content. To assist this process, there are a number of best practice guidelines publishers and users can follow.

In this module we look at the 5-Stars of linked open data and discover how this can be used to measure the technical usability of data.

In this module, we will explore the following:

- ✓ What are the 5-stars of linked open data
- ✓ The first three stars
- ✓ How to recognise the stars in data

Lesson 11 - How to clean your data

One of the biggest challenges when working with any data is errors. Often errors are not even noticed by data publishers because the data can change over many years. In other cases, errors can be the result of human mistakes in data entry, like mistyping or incorrect abbreviations.

When working with any data, it is important to know how to find errors and correct them to make the data more useful.

In this module, we will explore the following:

- ✓ Common data errors
- ✓ Useful data cleaning tools
- ✓ Why clean data?

Lesson 12 - Finding hidden data on the web

'Open data' does not only mean datasets available to download. Downloadable open data represents only a small fraction of the available data on the web.

The majority of data available on the web is hidden from the human eye. However, machines can find and read this data. In this module we look at techniques to unlock hidden data.

In this module, we will explore the following:

- ✓ How to locate hidden data
- ✓ What benefits hidden data can provide
- ✓ How to obtain hidden data

EUROPEAN
DATA PORTAL

Lesson 13 - Linking up the web of data

The current web is configured as a series of pages or 'documents'. While these documents draw on rich sources of data, they disguise it beneath pages designed for humans to view. In this module, we explore what would happen if all the pages or documents were removed from the web.

Imagine you only had the raw data, all open, all usable and all linked together in a network or 'web' of data. This module also introduces the web of open linked data and look at how the 5-stars of linked open data provide a roadmap for achieving this vision.

In this module, we will explore the following:

- ✓ What is the web of data?
- ✓ How web identifiers are used
- ✓ What a web of open linked data looks like

Lesson 14 - Open data advocacy

Opening data can be valuable for any organisation. Whether to drive innovation in the business, develop a clearer picture of operations or improve products and services, a growing number of private and public sector organisations now benefit from publishing and using open data.

However, opening data might require a change in the culture of an organisation. Most organisations are configured to protect their data resources, even when the benefits of openness outweigh the costs. The key to overcoming this resistance is clear: effective communication of the benefits that open data can bring.

In this module, we will explore the following:

- ✓ Structure the perfect pitch
- ✓ Identify different audiences for your pitch
- ✓ Tailor your pitch to different audiences

Lesson 15 - Managing change in open data

Change management is the process of introducing a long-term transition, change or development in an organisation. Change management is especially important when dealing with large changes, such as the cultural changes needed to succeed with open data.

In this module, we'll explore why it is important for organisations to manage change and key elements to help build a successful change management process.

In this module, we will explore the following:

- ✓ Structure the perfect pitch
- ✓ Identify different audiences for your pitch
- ✓ Tailor your pitch to different audiences

Lesson 16 - Achieving impact with open data

The goal of any open data project or initiative is to have some kind of impact, be it political transparency and accountability, social benefit or economic growth. Creating, finding and promoting examples of open data being used to improve our daily lives is one of the most challenging aspects of leading an open data project, programme or initiative.

In this module, we will explore the following:

- ✓ Structure the perfect pitch
- ✓ Identify different audiences for your pitch
- ✓ Tailor your pitch to different audiences

Further online material per topic

Open Government Data, PSI and the Directive

Author/Organisation	Link
Open Data Support	https://www.europeandataportal.eu/en/resources/more-training-materials
European Commission	https://ec.europa.eu/digital-agenda/en/implementation-public-sector-information-directive-member-states

Legislation and Licensing

The ODI	https://theodi.org/guides/publishers-guide-open-data-licensing
Open Data Support	https://www.europeandataportal.eu/en/resources/more-training-materials
ePSI Platform	http://www.europeandataportal.eu/sites/default/files/data-protection-in-re-use-psi.pdf
European Commission	https://ec.europa.eu/digital-agenda/en/implementation-public-sector-information-directive
European Commission	http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31995L0046

Open Data Lifecycle and Metadata

Open Data Support	http://www.europeandataportal.eu/en/content/training-library/library/training-materials
W3C Foundation	http://w3c.github.io/dwbp/bp.html#metadata

Linked Data, RDF, URIs, SPARQL

EUCLID (advanced)	http://www.euclid-project.eu/
Linked Data Tools	http://www.linkeddatatools.com/introducing-rdf
Open Data Support	https://www.europeandataportal.eu/en/resources/more-training-materials
ePSI Platform	http://www.europeandataportal.eu/sites/default/files/data-protection-in-re-use-psi.pdf
European Commission	https://joinup.ec.europa.eu/sites/default/files/c0/7d/10/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf
Fabien Gandon / Slideshare	http://www.slideshare.net/fabien_gandon/rdf-in-a-nutshell-v1

Vocabularies & Specifications

European Commission	https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final#download-links
The ODI	https://theodi.org/guides/marking-up-your-dataset-with-dcat
Open Data Support	https://www.europeandataportal.eu/en/resources/more-training-materials

EUROPEAN
DATA PORTAL

Training Companion

Do you want to provide training to your colleagues or others that want to learn more about Open Data? The topics covered by the eLearning modules are now available as an Open Data training guideline as well. The training collection offers materials and resources to deliver sessions on every aspect of Open Data. Simply choose a session plan, customise it with your content and deliver.

Discover the training material:

https://www.europeandataportal.eu/en/resources/training-companion