Open Data Handbook Documentation

Release 1.0.0

Open Knowledge Foundation

November 14, 2012

Contents

This handbook discusses the legal, social and technical aspects of open data. It can be used by anyone but is especially designed for those seeking to **open up** data. It discusses the **why, what and how** of open data – why to go open, what open is, and the how to 'open' data.

To get started, you may wish to look at the `Introduction`. You can navigate through the report using the Table of Contents (see sidebar or below).

We warmly welcome comments on the text and will incorporate feedback as we go forward. We also welcome contributions or suggestions for additional sections and areas to examine.

CHAPTER 1

Table of Contents

1.1 Introduction

Do you know exactly how much of your tax money is spent on street lights or on cancer research? What is the shortest, safest and most scenic bicycle route from your home to your work? And what is in the air that you breathe along the way? Where in your region will you find the best job opportunities and the highest number of fruit trees per capita? When can you influence decisions about topics you deeply care about, and whom should you talk to?

New technologies now make it possible to build the services to answer these questions automatically. Much of the data you would need to answer these questions is generated by public bodies. However, often the data required is not yet available in a form which is easy to use. This book is about how to unlock the potential of official and other information to enable new services, to improve the lives of citizens and to make government and society work better.

The notion of *open data* and specifically *open government data* - information, public or otherwise, which anyone is free to access and re-use for any purpose - has been around for some years. In 2009 open data started to become visible in the mainstream, with various governments (such as the USA, UK, Canada and New Zealand) announcing new initiatives towards opening up their public information.

This book explains the basic concepts of 'open data', especially in relation to government. It covers how open data creates value and can have a positive impact in many different areas. In addition to exploring the background, the handbook also provides concrete information on how to produce open data.

1.1.1 Target Audience

This handbook has a broad audience:

- for those who have never heard of open data before and those who consider themselves seasoned 'data professionals'
- for civil servants and for activists
- for journalists and researchers
- for politicians and developers
- for data geeks and those who have never heard of an API.

Most of the information currently provided is focused on data held by the public sector. However, the authors intentions are to broaden this as time permits. You are welcome to participate to help us with that effort.

This handbook is intended for those with little or no knowledge of the topic. If you do find a piece of jargon or terminology with which you aren't familiar, please see the detailed Glossary and FAQs (frequently asked questions) which can be found at the end of the handbook.

1.1.2 Credits

Credits and Copyright

Contributing authors

- Daniel Dietrich
- Jonathan Gray
- Tim McNamara
- Antti Poikola
- Rufus Pollock
- Julian Tait
- Ton Zijlstra

Existing sources directly used

- Technical Proposal for how IATI is implemented. *The IATI Technical Advisory Group led by Simon Parrish*
- Unlocking the Potential of Aid Information. *Rufus Pollock, Jonathan Gray, Simon Parrish, Jordan Hatcher*
- Finnish manual authored by *Antti Poikola*
- Beyond Access Report. *Access Info and the Open Knowledge Foundation*

Other sources

- W3C Publishing Government Data (2009) http://www.w3.org/TR/gov-data/

1.2 Why Open Data?

Open data, especially *open government data*, is a tremendous resource that is as yet largely untapped. Many individuals and organisations collect a broad range of different types of data in order to perform their tasks. Government is particularly significant in this respect, both because of the quantity and centrality of the data it collects, but also because most of that government data is public data by law, and therefore could be made open and made available for others to use. Why is that of interest?

There are many areas where we can expect open data to be of value, and where examples of how it has been used already exist. There are also many different groups of people and organisations who can benefit from the availability of open data, including government itself. At the same time it is impossible to predict precisely how and where value will be created in the future. The nature of innovation is that developments often comes from unlikely places.

It is already possible to point to a large number of areas where open government data is creating value. Some of these areas include:

- Transparency and democratic control
- Participation
- Self-empowerment
- Improved or new private products and services
- Innovation
- Improved efficiency of government services

- Improved effectiveness of government services
- Impact measurement of policies
- New knowledge from combined data sources and patterns in large data volumes

Examples exist for most of these areas.

In terms of transparency, projects such as the Finnish 'tax tree' and British 'where does my money go' show how your tax money is being spent by the government. And there's the example of how open data saved Canada $3.2 billion in charity tax fraud. Also various websites such as the Danish folketsting.dk track activity in parliament and the law making processes, so you can see what exactly is happening, and which parliamentarians are involved.

Open government data can also help you to make better decisions in your own life, or enable you to be more active in society. A woman in Denmark built findtoilet.dk, which showed all the Danish public toilets, so that people she knew with bladder problems can now trust themselves to go out more again. In the Netherlands a service, vervuilingsalarm.nl, is available which warns you with a message if the air-quality in your vicinity is going to reach a self-defined threshold tomorrow. In New York you can easily find out where you can walk your dog, as well as find other people who use the same parks. Services like 'mapumental' in the UK and 'mapnificent' in Germany allow you to find places to live, taking into account the duration of your commute to work, housing prices, and how beautiful an area is. All these examples use open government data.

Economically, open data is of great importance as well. Several studies have estimated the economic value of open data at several tens of billions of Euros annually in the EU alone. New products and companies are re-using open data. The Danish husetsweb.dk helps you to find ways of improving the energy efficiency of your home, including financial planning and finding builders who can do the work. It is based on re-using cadastral information and information about government subsidies, as well as the local trade register. Google Translate uses the enormous volume of EU documents that appear in all European languages to train the translation algorithms, thus improving its quality of service.

Open data is also of value for government itself. For example, it can increase government efficiency. The Dutch Ministry of Education has published all of their education-related data online for re-use. Since then, the number of questions they receive has dropped, reducing work-load and costs, and the remaining questions are now also easier for civil servants to answer, because it is clear where the relevant data can be found. Open data is also making government more effective, which ultimately also reduces costs. The Dutch department for cultural heritage is actively releasing their data and collaborating with amateur historical societies and groups such as the Wikimedia Foundation in order to execute their own tasks more effectively. This not only results in improvements to the quality of their data, but will also ultimately make the department smaller.

While there are numerous instances of the ways in which open data is already creating both social and economic value, we don't yet know what new things will become possible. New combinations of data can create new knowledge and insights, which can lead to whole new fields of application. We have seen this in the past, for example when Dr. Snow discovered the relationship between drinking water pollution and cholera in London in the 19th century, by combining data about cholera deaths with the location of water wells. This led to the building of London's sewage systems, and hugely improved the general health of the population. We are likely to see such developments happening again as unexpected insights flow from the combination of different open data sets.

This untapped potential can be unleashed if we turn public government data into open data. This will only happen, however, if it is really open, i.e. if there are no restrictions (legal, financial or technological) to its re-use by others. Every restriction will exclude people from re-using the public data, and make it harder to find valuable ways of doing that. For the potential to be realized, public data needs to be open data.

1.3 What is Open Data?

This handbook is about *open data* but what exactly is it? In particular what makes *open* data open, and what sorts of data are we talking about?

1.3.1 What is Open?

This handbook is about open data - but what exactly is *open* data? For our purposes, open data is as defined by the Open Definition:

Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.

The full Open Definition gives precise details as to what this means. To summarize the most important:

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Re-use and Redistribution:** the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

If you're wondering why it is so important to be clear about what open means and why this definition is used, there's a simple answer: **interoperability.**

Interoperability denotes the ability of diverse systems and organizations to work together (inter-operate). In this case, it is the ability to interoperate - or intermix - different datasets.

Interoperability is important because it allows for different components to work together. This ability to componentize and to 'plug together' components is essential to building large, complex systems. Without interoperability this becomes near impossible — as evidenced in the most famous myth of the Tower of Babel where the (in)ability to communicate (to interoperate) resulted in the complete breakdown of the tower-building effort.

We face a similar situation with regard to data. The core of a "commons" of data (or code) is that one piece of "open" material contained therein can be freely intermixed with other "open" material. This interoperability is absolutely key to realizing the main practical benefits of "openness": the dramatically enhanced ability to combine different datasets together and thereby to develop more and better products and services (these benefits are discussed in more detail in the section on 'why' open data).

Providing a clear definition of openness ensures that when you get two open datasets from two different sources, you will be able to combine them together, and it ensures that we **avoid our own 'tower of babel': lots of datasets but little or no ability to combine them together into the larger systems where the real value lies.**

1.3.2 What Data are You Talking About?

Readers have already seen examples of the sorts of data that are or may become open - and they will see more examples below. However, it will be useful to quickly outline what sorts of data are, or could be, open – and, equally importantly, what won't be open.

The key point is that when opening up data, the focus is on non-personal data, that is, data which does not contain information about specific individuals.

Similarly, for some kinds of government data, national security restrictions may apply.

1.4 How to Open up Data

This section forms the core of this handbook. It gives concrete, detailed advice on how data holders can open up data. We'll go through the basics, but also cover the pitfalls. Lastly, we will discuss the more subtle issues that can arise.

There are three key rules we recommend following when opening up data:

- **Keep it simple.** Start out small, simple and fast. There is no requirement that every dataset must be made open right now. Starting out by opening up just one dataset, or even one part of a large dataset, is fine – of course, the more datasets you can open up the better.

  Remember this is about innovation. Moving as rapidly as possible is good because it means you can build momentum and learn from experience – innovation is as much about failure as success and not every dataset will be useful.

- **Engage early and engage often.** Engage with actual and potential users and re-users of the data as early and as often as you can, be they citizens, businesses or developers. This will ensure that the next iteration of your service is as relevant as it can be.

  It is essential to bear in mind that much of the data will not reach ultimate users directly, but rather via 'info-mediaries'. These are the people who take the data and transform or remix it to be presented. For example, most of us don't want or need a large database of GPS coordinates, we would much prefer a map. Thus, engage with infomediaries first. They will re-use and repurpose the material.

- **Address common fears and misunderstandings**. This is especially important if you are working with or within large institutions such as government. When opening up data you will encounter plenty of questions and fears. It is important to (a) identify the most important ones and (b) address them at as early a stage as possible.

There are four main steps in making data open, each of which will be covered in detail below. These are in very approximate order - many of the steps can be done simultaneously.

1. **Choose your dataset(s).** Choose the dataset(s) you plan to make open. Keep in mind that you can (and may need to) return to this step if you encounter problems at a later stage.

2. **Apply an open license.**

   (a) Determine what intellectual property rights exist in the data.

   (b) Apply a suitable 'open' license that licenses all of these rights and supports the definition of openness discussed in the section above on 'What Open Data'

   (c) NB: if you can't do this go back to step 1 and try a different dataset.

3. **Make the data available** - in bulk and in a useful format. You may also wish to consider alternative ways of making it available such as via an API.

4. **Make it discoverable** - post on the web and perhaps organize a central catalog to list your open datasets.

1.4.1 Choose Dataset(s)

Choosing the dataset(s) you plan to make open is the first step – though remember that the whole process of opening up data is iterative and you can return to this step if you encounter problems later on.

If you already know exactly what dataset(s) you plan to open up you can move straight on to the next section. However, in many cases, especially for large institutions, choosing which datasets to focus on is a challenge. How should one proceed in this case?

Creating this list should be a quick process that identifies which datasets could be made open to start with. There will be time at later stages to check in detail whether each dataset is suitable.

There is **no requirement** to create a comprehensive list of your datasets. The main point to bear in mind is whether it is feasible to publish this data at all (whether openly or otherwise) - see *this previous section*.

Asking the community

We recommend that you ask the community in the first instance. That is the people who will be accessing and using the data, as they are likely to have a good understanding of which data could be valuable.

1. Prepare a short list of potential datasets that you would like feedback on. It is not essential that this list concurs with your expectations, the main intention is to get a feel for the demand. This could be based on other countries' *open data* catalogs.

2. Create a request for comment.
3. Publicise your request with a webpage. Make sure that it is possible to access the request through its own URL. That way, when shared via social media, the request can be easily found.
4. Provide easy ways to submit responses. Avoid requiring registration, as it reduces the number of responses.
5. Circulate the request to relevant mailing lists, forums and individuals, pointing back to the main webpage.
6. Run a consultation event. Make sure you run it at a convenient time where the average business person, data wrangler and official can attend.
7. Ask a politician to speak on your agency's behalf. Open data is very likely to be part of a wider policy of increasing access to government information.

Cost basis

How much money do agencies spend on the collection and maintainence of data that they hold? If they spend a great deal on a particular set of data, then it is highly likely that others would like to access it.

This argument may be fairly susceptible to concerns of freeriding. The question you will need to respond to is, "Why should other people get information for free that is so expensive?". The answer is that the expense is absorbed by the public sector to perform a particular function. The cost of sending that data, once it has been collected, to a third party is approximately nothing. Therefore, they should be charged nothing.

Ease of release

Sometimes, rather than deciding which data would be most valuable, it could be useful to take a look at which data is easiest to get into the public's hands. Small, easy releases can act as the catalyst for larger behavioural change within organisations.

Be careful with this approach however. It may be the case that these small releases are of so little value that nothing is built from them. If this occurs, faith in the entire project could be undermined.

Observe peers

Open data is a growing movement. There are likely to be many people in your area who understand what other areas are doing. Formulate a list on the basis of what those agencies are doing.

1.4.2 Apply an Open License (Legal Openness)

In most jurisdictions there are intellectual property rights in data that prevent third-parties from using, reusing and redistributing data without explicit permission. Even in places where the existence of rights is uncertain, it is important to apply a license simply for the sake of clarity. Thus, **if you are planning to make your data available you should put a license on it** – and if you want your data to be open this is even more important.

What licenses can you use? We recommend that for 'open' data you use one of the licenses conformant with the Open Definition and marked as suitable for data. This list (along with instructions for usage) can be found at:

- http://opendefinition.org/licenses/

A short 1-page instruction guide to applying an open data license can be found on the Open Data Commons site:

- http://opendatacommons.org/guide/

1.4.3 Make Data Available (Technical Openness)

Open data needs to be technically open as well as legally open. Specifically, the data needs to be available in bulk in a *machine-readable* format.

Available Data should be priced at no more than a reasonable cost of reproduction, preferably as a free download from the Internet. This pricing model is achieved because your agency should not undertake any cost when it provides data for use.

In bulk The data should be available as a complete set. If you have a register which is collected under statute, the entire register should be available for download. A web API or similar service may also be very useful, but they are not a substitutes for bulk access.

In an open, machine-readable format Re-use of data held by the public sector should not be subject to patent restrictions. More importantly, making sure that you are providing machine-readable formats allows for greatest re-use. To illustrate this, consider statistics published as PDF (Portable Document Format) documents, often used for high quality printing. While these statistics can be read by humans, they are very hard for a computer to use. This greatly limits the ability for others to re-use that data.

Here are a few policies that will be of great benefit:

- Keep it simple,
- Move fast
- Be pragmatic.

In particular it is better to give out raw data now than perfect data in six months' time.

There are many different ways to make data available to others. The most natural in the Internet age is online publication. There are many variations to this model. At its most basic, agencies make their data available via their websites and a central catalog directs visitors to the appropriate source. However, there are alternatives.

When *connectivity* is limited or the size of the data extremely large, distribution via other formats can be warranted. This section will also discuss alternatives, which can act to keep prices very low.

Online methods

Via your existing website

The system which will be most familiar to your web content team is to provide files for download from webpages. Just as you currently provide access to discussion documents, data files are perfectly happy to be made available this way.

One difficulty with this approach is that it is very difficult for an outsider to discover where to find updated information. This option places some burden on the people creating tools with your data.

Via 3rd party sites

Many repositories have become hubs of data in particular fields. For example, pachube.com is designed to connect people with sensors to those who wish to access data from them. Sites like Infochimps.com and Talis.com allow public sector agencies to store massive quantities of data for free.

Third party sites can be very useful. The main reason for this is that they have already pooled together a community of interested people and other sets of data. When your data is part of these platforms, a type of positive compound interest is created.

Wholesale data platforms already provide the infrastructure which can support the demand. They often provide analytics and usage information. For public sector agencies, they are generally free.

These platforms can have two costs. The first is independence. Your agency needs to be able to yield control to others. This is often politically, legally or operationally difficult. The second cost may be openness. Ensure that your data platform is agnostic about who can access it. Software developers and scientists use many operating sytems, from smart phones to supercomputers. They should all be able to access the data.

Via FTP servers

A less fashionable method for providing access to files is via the File Transfer Protocol (FTP). This may be suitable if your audience is technical, such as software developers and scientists. The FTP system works in place of HTTP, but is specifically designed to support file transfers.

FTP has fallen out of favour. Rather than providing a website, looking through an FTP server is much like looking through folders on a computer. Therefore, even though it is fit for purpose, there is far less capacity for web development firms to charge for customisation.

As torrents

BitTorrent is a system which has become familiar to policy makers because of its association with copyright infringement. BitTorrent uses files called torrents, which work by splitting the cost of distributing files between all of the people accessing those files. Instead of servers becoming overloaded, the supply increases with the demand increases. This is the reason that this system is so successful for sharing movies. It is a wonderfully efficient way to distribute very large volumes of data.

As an API

Data can be published via an *Application Programming Interface* (API). These interfaces have become very popular. They allow programmers to select specific portions of the data, rather than providing all of the data in bulk as a large file. APIs are typically connected to a database which is being updated in real-time. This means that making information available via an API can ensure that it is up to date.

Publishing raw data in bulk should be the primary concern of all open data intiatives. There are a number of costs to providing an API:

1. The price. They require much more development and maintainence than providing files.
2. The expectations. In order to foster a community of users behind the system, it is important to provide certainty. When things go wrong, you will be expected to incur the costs of fixing them.

Access to bulk data ensures that:

1. there is no dependency on the original provider of the data, meaning that if a restructure or budget cycle changes the situation, the data are still available.
2. anyone else can obtain a copy and redistribute it. This reduces the cost of distribution away from the source agency and means that there is no single point of failure.
3. others can develop their own services using the data, because they have certainty that the data will not be taken away from them.

Providing data in bulk allows others to use the data beyond its original purposes. For example, it allows it to be converted into a new format, linked with other resources, or versioned and archived in multiple places. While the latest version of the data may be made available via an API, raw data should be made available in bulk at regular intervals.

For example, the Eurostat statistical service has a bulk download facility offering over 4000 data files. It is updated twice a day, offers data in *Tab-separated values* (TSV) format, and includes documentation about the download facility as well as about the data files.

Another example is the District of Columbia Data Catalog, which allows data to be downloaded in CSV and XLS format in addition to live feeds of the data.

1.4.4 Make data discoverable

Open data is nothing without users. You need to be able to make sure that people can find the source material. This section will cover different approaches.

The most important thing is to provide a neutral space which can overcome both inter-agency politics and future budget cycles. Jurisdictional borders, whether sectorial or geographical, can make cooperation difficult. However, there are significant benefits in joining forces. The easier it is for outsiders to discover data, the faster new and useful tools will be built.

Existing tools

There are a number of tools which are live on the web that are specifically designed to make data more discoverable.

One of the most prominent is the DataHub and is a catalog and data store for datasets from around the world. The site makes it easy for individuals and organizations to publish material and for data users to find material they need.

In addition, there are dozens of specialist catalogs for different sectors and places. Many scientific communities have created a catalog system for their fields, as data are often required for publication.

For government

As it has emerged, orthodox practice is for a lead agency to create a catalog for the government's data. When establishing a catalog, try to create some structure which allows many departments to easily keep their own information current.

Resist the urge to build the software to support the catalog from scratch. There are free and open source software solutions (such as CKAN) which have been adopted by many governments already. As such, investing in another platform may not be needed.

There are a few things that most open data catalogs miss. Your programme could consider the following:

- Providing an avenue to allow the private and community sectors to add their data. It may be worthwhile to think of the catalog as the region's catalog, rather than the regional government's.
- Facilitating improvement of the data by allowing derivatives of datasets to be cataloged. For example, someone may geocode addresses and may wish to share those results with everybody. If you only allow single versions of datasets, these improvements remain hidden.
- Be tolerant of your data appearing elsewhere. That is, content is likely to be duplicated to communities of interest. If you have river level monitoring data available, then your data may appear in a catalog for hydrologists.
- Ensure that access is equitable. Try to avoid creating a privileged level of access for officials or tenured researchers as this will undermine community participation and engagement.

For civil society

Be willing to create a supplementary catalog for non-official data.

It is very rare for governments to associate with unofficial or non-authoritative sources. Officials have often gone to great expense to ensure that there will not be political embarrassment or other harm caused from misuse or overreliance on data.

Moreover, governments are unlikely to be willing to support activities that mesh their information with information from businesses. Governments are rightfully skeptical of profit motives. Therefore, an independent catalog for community groups, businesses and others may be warranted.

1.5 So I've Opened Up Some Data, Now What?

We've looked at how to make government information legally and technically reusable. The next step is to encourage others to make use of that data.

This section looks at additional things which can be done to promote data re-use.

1.5.1 Tell the world!

First and foremost, make sure that you promote the fact that you've embarked on a campaign to promote *open data* in your area of responsibility.

If you open up a bunch of datasets, it's definitely worth spending a bit of time to make sure that people know (or at least can find out) that you've done so.

In addition to things like press releases, announcements on your website, and so on, you may consider:

- Contacting prominent organisations or individuals who work/are interested in this area
- Contacting relevant mailing lists or social networking groups
- Directly contacting prospective users who you know may be interested in this data

Understanding your audience

Like all public communication, engaging with the data community needs to be targeted. Like all stakeholder groups, the right message can be wasted if it is directed to the wrong area.

Digital communities tend to be very willing to share new information, yet they very rapidly consume it. Write as if your messages will be skimmed over, rather than critically examined in-depth.

Members of the tech community are less likely than the general public to use MS Windows. This means that you should not save documents in MS Office formats which can be read offline. There are two resons for this:

- The first is that those documents will be less accessible. Rather than the document you see on your screen, readers may see an imperfect copy from an alternative.
- Secondly, your agency sends an implicit message that you are unwilling to take a step towards developers. Instead, you show that you are expecting the technology community to come to you.

Post your material on third-party sites

Many blogs have created a large readership in specialised topic areas. It may be worthwhile adding an article about your initiative on their site. These can be mutually beneficial. You receive more interest and they receive a free blog post in their topic area.

Making your communications more social-media friendly

It's unrealistic to expect that officials should spend long periods of time engaging with social media. However, there are several things that you can do to make sure that your content can be easily shared between technical users. Some tips:

Provide unique pages for each piece of content When a message is shared with others, the recipient of the referral will be looking for the relevant content quickly.

Avoid making people download your press releases Press releases are fine. They are concise messages about a particular point. However, if you require people to download the content and for it to open outside of a web browser, then fewer people will read it. Search engines are less likely to index the content. People are less likely to click to download.

Consider using an Open license for your content Apart from providing certainty to people who wish to share your content that this is permissible, you send a message that your agency understands openness. This is bound to leave an impression far more significant to proponents of open data than any specific sentence in your press release.

Social media

It’s inefficient for cash-strapped agencies to spend hours on social media sites. The most significant way that your voice can be heard through these fora is by making sure that blog posts are easily shareable. That means, before reading the next section, make sure that you have read the last. With that in mind, here are a few suggestions:

Discussion fora Twitter has emerged as the platform of choice for disseminating information rapidly. Anything tagged with #opendata will be immediately seen by thousands.

LinkedIn has a large selection of groups which are targeted towards open data.

While Facebook is excellent for a general audience, it has not received a great deal of attention in the open data community.

Link aggregators Submit your content to the equivalent of newswires for geeks. Reddit and Hacker News are the two biggest in this arena at the moment. To a lesser extent, Slashdot and Digg are also useful tools in this area.

These sites have a tendency to drive significant traffic to interesting material. They are also heavily focused on topic areas.

1.5.2 Getting folks in a room: Unconferences, Meetups and Barcamps

Face-to-face events can be a very effective way to encourage others to use your data. Reasons that you may consider putting on an event include:

- Finding out more about prospective re-users
- Finding out more about demand for different datasets
- Finding out more about how people want to re-use your data
- Enabling prospective re-users to find out more about what data you have
- Enabling prospective users to meet each other (e.g. so they can collaborate)
- Exposing your data to a wider audience (e.g. from blog posts or media coverage that the event may help to generate)

There are also lots of different ways of running events, and different types of events, depending on what aim you want to achieve. As well as more traditional conference models, which will include things like preprepared formal talks, presentations and demonstrations, there are also various kinds of participant driven events, where those who turn up may:

- Guide or define the agenda for the event
- Introduce themselves, talk about what they’re interested in and what they’re working on, on an ad hoc basis
- Give impromptu micro-short presentations on something they are working on
- Lead sessions on something they are interested in

There is plenty of documentation online about how to run these kinds of events, which you can find by searching for things like: ‘unconference’, ‘barcamp’, ‘meetup’, ‘speedgeek’, ‘lightning talk’, and so on. You may also find it worthwhile to contact people who have run these kinds of events in other countries, who will most likely be keen to help you out and to advise you on your event. It may be valuable to partner with another organisation (e.g. a civic society organisation, a news organisation or an educational institution) to broaden your base participants and to increase your exposure.

1.5.3 Making things! Hackdays, prizes and prototypes

The structure of these competitions is that a number of datasets are released and programmers then have a short time-frame - running from as little as 48 hours to a few weeks - to develop applications using the data. A prize is then awarded to the best application. Competitions have been held in a number of countries including the UK, the US, Norway, Australia, Spain, Denmark and Finland.

Examples for Competitions

Show us a better way was the first such competition in the world. It was initiated by the UK Government's "The Power of Information Taskforce" headed by Cabinet Office Minister Tom Watson in March 2008. This competition asked "What would you create with public information?" and was open to programmers from around the world, with a tempting £80,000 prize for the five best applications.

Apps for Democracy, one of the first competitions in the United States, was launched in October 2008 by Vivek Kundra, at the time Chief Technology Officer (CTO) of the District of Columbia (DC) Government. Kundra had developed the groundbreaking DC data catalog, http://data.octo.dc.gov/, which included datasets such as real-time crime feeds, school test scores, and poverty indicators. It was at the time the most comprehensive local data catalog in the world. The challenge was to make it useful for citizens, visitors, businesses and government agencies of Washington, DC.

The creative solution was to create the Apps for Democracy contest. The strategy was to ask people to build applications using the data from the freshly launched data catalog. It included an online submission for applications, many small prizes rather than a few large ones, and several different categories as well as a "People's Choice" prize. The competition was open for 30 days and cost the DC government $50,000. In return, a total of 47 iPhone, Facebook and web applications were developed with an estimated value in excess of $2,600,000 for the local economy.

The Abre Datos (Open Data) Challenge 2010. Held in Spain in April 2010, this contest invited developers to create open source applications making use of public data in just 48 hours. The competition had 29 teams of participants who developed applications that included a mobile phone programme for accessing traffic information in the Basque Country, and for accessing data on buses and bus stops in Madrid, which won the first and second prizes of €3,000 and €2,000 respectively.

Nettskap 2.0. In April 2010 the Norwegian Ministry for Government Administration held "Nettskap 2.0". Norwegian developers – companies, public agencies or individuals – were challenged to come up with web-based project ideas in the areas of service development, efficient work processes, and increased democratic participation. The use of government data was explicitly encouraged. Though the application deadline was just a month later, on May 9, the Minister Rigmor Aasrud said the response was "overwhelming". In total 137 applications were received, no less than 90 of which built on the re-use of government data. A total amount of NOK 2.5 million was distributed among the 17 winners; while the total amount applied for by the 137 applications was NOK 28.4 million.

Mashup Australia. The Australian Government 2.0 Taskforce invited citizens to show why open access to Australian government information would be positive for the country's economy and social development. The contest ran from October 7th to November 13th 2009. The Taskforce released some datasets under an open license and in a range of reusable formats. The 82 applications that were entered into the contest are further evidence of the new and innovative applications which can result from releasing government data on open terms.

Conferences, Barcamps, Hackdays

One of the more effective ways for Civil Society Organisations (CSOs) to demonstrate to governments the value of opening up their datasets is to show the multiple ways in which the information can be managed to achieve social and economic benefit. CSOs that promote re-use have been instrumental in countries where there have been advances in policy and law to ensure that datasets are both technically and legally open.

The typical activities which are undertaken as part of these initiatives normally include competitions, *open government data* conferences, "unconferences", workshops and "hack days". These activities are often organised by the user community with data that has already been published proactively or obtained using access to information requests. In other cases, civil society advocates have worked with progressive public officials to secure new release of datasets that can be used by programmers to create innovative applications.

1.6 Glossary

Anonymisation The process of adapting data so that individuals cannot be identified from it.

Anonymization See *Anonymisation*.

API See *Application Programming Interface*.

Application Programming Interface A way computer programs talk to one another. Can be understood in terms of how a programmer sends instructions between programs.

AR See *Information Asset Register*.

Attribution License A license that requires that the original source of the licensed material is cited (attributed).

BitTorrent BitTorrent is a protocol for distributing the bandwith for transferring very large files between the computers which are participating in the transfer. Rather than downloading a file from a specific source, BitTorrent allows peers to download from each other.

Connectivity Connectivity relates to the ability for communities to connect to the Internet, especially the World Wide Web.

Copyright A right for the creators of creative works to restrict others' use of those works. An owner of copyright is entitled to determine how others may use that work.

DAP See *Data Access Protocol*.

Data Access Protocol A system that allows outsiders to be granted access to databases without overloading either system.

Data protection legislation Data protection legislation is not about protecting the data, but about protecting the right of citizens to live without fear that information about their private lives might become public. The law protects privacy (such as information about a person's economic status, health and political position) and other rights such as the right to freedom of movement and assembly. For example, in Finland a travel card system was used to record all instances when the card was shown to the reader machine on different public transport lines. This raised a debate from the perspective of freedom of movement and the travel card data collection was abandoned based on the data protection legislation.

Database rights A right to prevent others from extracting and reusing content from a database. Exists mainly in European jurisdictions.

EU European Union.

EU PSI Directive The *Directive on the re-use of public sector information*, 2003/98/EC. "deals with the way public sector bodies should enhance re-use of their information resources." Legislative Actions - PSI Directive

IAR See *Information Asset Register*.

Information Asset Register IARs are registers specifically set up to capture and organise meta-data about the vast quantities of information held by government departments and agencies. A comprehensive IAR includes databases, old sets of files, recent electronic files, collections of statistics, research and so forth.

The *EU PSI Directive* recognises the importance of asset registers for prospective re-users of public information. It requires member states to provide lists, portals, or something similar. It states:

```
Tools that help potential re-users to find documents available
for re-use and the conditions for re-use can facilitate
considerably the cross-border use of public sector documents.
Member States should therefore ensure that practical arrangements
are in place that help re-users in their search for documents
available for reuse. Assets lists, accessible preferably online,
of main documents (documents that are extensively re-used or
that have the potential to be extensively re-used), and portal
sites that are linked to decentralised assets lists are examples
of such practical arrangements.
```

IARs can be developed in different ways. Government departments can develop their own IARs and these can be linked to national IARs. IARs can include information which is held by public bodies but which has not yet been – and maybe will not be – proactively published. Hence they allow members of the public to identify information which exists and which can be requested.

For the public to make use of these IARs, it is important that any registers of information held should be as complete as possible in order to be able to have confidence that documents can be found. The incompleteness of some registers is a significant problem as it creates a degree of unreliability which may discourage some from using the registers to search for information.

It is essential that the metadata in the IARs should be comprehensive so that search engines can function effectively. In the spirit of open government data, public bodies should make their IARs available to the general public as raw data under an open license so that civic hackers can make use of the data, for example by building search engines and user interfaces.

Intellectual property rights Monopolies granted to individuals for intellectual creations.

IP rights See *Intellectual property rights*.

Machine-readable Formats that are machine readable are ones which are able to have their data extracted by computer programs easily. PDF documents are not machine readable. Computers can display the text nicely, but have great difficulty understanding the context that surrounds the text.

Open Data Open data are able to be used for any purpose. More details can be read at opendefinition.org.

Open Government Data *Open data* produced by the government. This is generally accepted to be data gathered during the course of business as usual activities which do not identify individuals or breach commercial sensitivity. Open government data is a subset of *Public Sector Information*, which is broader in scope. See http://opengovernmentdata.org for details.

Open standards Generally understood as technical standards which are free from licencing restrictions. Can also be interpreted to mean standards which are developed in a vendor-neutral manner.

PSI See *Public Sector Information*.

Public domain No copyright exists over the work. Does not exist in all jurisdictions.

Public Sector Information Information collected or controlled by the public sector.

Re-use Use of content outside of its original intention.

Share-alike License A license that requires users of a work to provide the content under the same or similar conditions as the original.

Tab-separated values Tab-separated values (TSV) are a very common form of text file format for sharing tabular data. The format is extremely simple and highly *machine-readable*.

Web API An *API* that is designed to work over the Internet.

1.7 Appendices

1.7.1 File Formats

An Overview of File Formats

JSON

JSON is a simple file format that is very easy for any programming language to read. Its simplicity means that it is generally easier for computers to process than others, such as XML.

XML

XML is a widely used format for data exchange because it gives good opportunities to keep the structure in the data and the way files are built on, and allows developers to write parts of the documentation in with the data without interfering with the reading of them.

RDF

A W3C-recommended format called RDF makes it possible to represent data in a form that makes it easier to combine data from multiple sources. RDF data can be stored in XML and JSON, among other serializations. RDF encourages the use of URLs as identifiers, which provides a convenient way to directly interconnect existing *open data* initiatives on the Web. RDF is still not widespread, but it has been a trend among Open Government initiatives, including the British and Spanish Government Linked Open Data projects. The inventor of the Web, Tim Berners-Lee, has recently proposed a five-star scheme that includes linked RDF data as a goal to be sought for open data initiatives.

Spreadsheets

Many authorities have information left in the spreadsheet, for example Microsoft Excel. This data can often be used immediately with the correct descriptions of what the different columns mean.

However, in some cases there can be macros and formulas in spreadsheets, which may be somewhat more cumbersome to handle. It is therefore advisable to document such calculations next to the spreadsheet, since it is generally more accessible for users to read.

Comma Separated Files

CSV files can be a very useful format because it is compact and thus suitable to transfer large sets of data with the same structure. However, the format is so spartan that data are often useless without documentation since it can be almost impossible to guess the significance of the different columns. It is therefore particularly important for the comma-separated formats that documentation of the individual fields are accurate.

Furthermore it is essential that the structure of the file is respected, as a single omission of a field may disturb the reading of all remaining data in the file without any real opportunity to rectify it, because it cannot be determined how the remaining data should be interpreted.

Text Document

Classic documents in formats like Word, ODF, OOXML, or PDF may be sufficient to show certain kinds of data - for example, relatively stable mailing lists or equivalent. It may be cheap to exhibit in, as often it is the format the data is born in. The format gives no support to keep the structure consistent, which often means that it is difficult to enter data by automated means. Be sure to use templates as the basis of documents that will display data for re-use, so it is at least possible to pull information out of documents.

It can also support the further use of data to use typography markup as much as possible so that it becomes easier for a machine to distinguish headings (any type specified) from the content and so on. Generally it is recommended not to exhibit in word processing format, if data exists in a different format.

Plain Text

Plain text documents (.txt) are very easy for computers to read. They generally exclude structural metadata from inside the document however, meaning that developers will need to create a parser that can interpret each document as it appears.

Some problems can be caused by switching plain text files between operating systems. MS Windows, Mac OS X and other Unix variants have their own way of telling the computer that they have reached the end of the line.

Scanned image

Probably the least suitable form for most data, but both TIFF and JPEG-2000 can at least mark them with documentation of what is in the picture - right up to mark up an image of a document with full text content of the document. It may be relevant to their displaying data as images whose data are not born electronically - an obvious example is the old church records and other archival material - and a picture is better than nothing.

Proprietary formats

Some dedicated systems, etc. have their own data formats that they can save or export data in. It can sometimes be enough to expose data in such a format - especially if it is expected that further use would be in a similar system as that which they come from. Where further information on these proprietary formats can be found should always be indicated, for example by providing a link to the supplier's website. Generally it is recommended to display data in non-proprietary formats where feasible.

HTML

Nowadays much data is available in HTML format on various sites. This may well be sufficient if the data is very stable and limited in scope. In some cases, it could be preferable to have data in a form easier to download and manipulate, but as it is cheap and easy to refer to a page on a website, it might be a good starting point in the display of data.

Typically, it would be most appropriate to use tables in HTML documents to hold data, and then it is important that the various data fields are displayed and are given IDs which make it easy to find and manipulate data. Yahoo has developed a tool (http://developer.yahoo.com/yql/) that can extract structured information from a website, and such tools can do much more with the data if it is carefully tagged.

Open File Formats

Even if information is provided in electronic, machine-readable format, and in detail, there may be issues relating to the format of the file itself.

The formats in which information is published – in other words, the digital base in which the information is stored - can either be "open" or "closed". An open format is one where the specifications for the software are available to anyone, free of charge, so that anyone can use these specifications in their own software without any limitations on re-use imposed by intellectual property rights.

If a file format is "closed", this may be either because the file format is proprietary and the specification is not publicly available, or because the file format is proprietary and even though the specification has been made public, re-use is limited. If information is released in a closed file format, this can cause significant obstacles to reusing the information encoded in it, forcing those who wish to use the information to buy the necessary software.

The benefit of open file formats is that they permit developers to produce multiple software packages and services using these formats. This then minimises the obstacles to reusing the information they contain.

Using proprietary file formats for which the specification is not publicly available can create dependence on third-party software or file format license holders. In worst-case scenarios, this can mean that information can only be read using certain software packages, which can be prohibitively expensive, or which may become obsolete.

The preference from the *open government data* perspective therefore is that information be released in **open file formats which are machine-readable.**

Example: UK traffic data

Andrew Nicolson is a software developer who was involved in an (ultimately successful) campaign against the construction of a new road, the Westbury Eastern bypass, in the UK. Andrew was interested in accessing and using the road traffic data that was being used to justify the proposals. He managed to obtain some of the relevant data via freedom of information requests, but the local government provided the data in a proprietary format which can only be read using software produced by a company called Saturn, who specialise in traffic modelling and forecasting. There is no provision for a "read only" version of the software, so Andrew's group had no choice but to purchase a software license, eventually paying £500 (€600) when making use of an educational discount. The main software packages on the April 2010 price list from Saturn start at £13,000 (over €15,000), a price which is beyond the reach of most ordinary citizens.

Although no access to information law gives a right of access to information in open formats, open government data initiatives are starting to be accompanied by policy documents which stipulate that official information must be made available in open file formats. Setting the gold standard has been the Obama Administration, with the Open Government Directive issued in December 2009, which says:

> *To the extent practicable and subject to valid restrictions, agencies should publish information online in an open format that can be retrieved, downloaded, indexed, and searched by commonly used web search applications. An open format is one that is platform independent, machine readable, and made available to the public without restrictions that would impede the re-use of that information.*

How do I use a given format?

When an authority must exhibit new data – data that has not been exhibited before – you should choose the format that provides the best balance between cost and suitability for purpose. For each format there are some things you should be aware of, and this section aims to explain them.

This section focuses only on how the cut surfaces are best arranged so that machines can access them directly. Advice and guidance about how web sites and web solutions should be designed can be found elsewhere.

Web services

For data that changes frequently, and where each pull is limited in size, it is very relevant to expose data through web services. There are several ways to create a web service, but some of the most used is SOAP and REST. Generally, SOAP over REST, REST services, but are very easy to develop and use, so it is a widely used standard.

Database

Like web services, databases provide direct access to data dynamically. Databases have the advantage that they can allow users to put together just the extraction that they are interested in.

There are some security concerns about allowing remote database extraction and database access is only useful if the structure of the database and the importance of individual tables and fields are well documented. Often, it is relatively simple and inexpensive to create web services that expose data from a database, which can be an easy way to address safety concerns.

1.7.2 What Legal (IP) Rights Are There in Data(bases)

When talking about data(bases) we first need to distinguish between the structure and the content of a database (when we use the term 'data' we shall mean the content of the database itself). Structural elements include things like the field names and a model for the data – the organization of these fields and their inter-relation.

In many jurisdictions it is likely that the structural elements of a database will be covered by copyright (it depends somewhat on the level of 'creativity' involved in creating this structure).

However, here we are particularly interested in the data. When we talk of "data" we need to be a bit careful because the word isn't particularly precise: "data" can mean a few items or even a single item (for example a single bibliographic record, a lat/long etc) or "data" can mean a large collection (e.g. all the material in the database). To avoid confusion we shall reserve the term "content" to mean the individual items, and data to denote the collection.

Unlike for material such as text, music or film, the legal situation for data varies widely across countries. However, most jurisdictions **do** grant some rights in the data (as a collection).

This distinction between the "content" of a database and the collection is especially crucial for factual databases since no jurisdiction grants a monopoly right over the individual facts (the "content"), even though it may grant right(s) in them as a collection. To illustrate, consider the simple example of a database which lists the melting point of various substances. While the database as a whole might be protected by law so that one is not allow to access, re-use or redistribute it without permission, this would never prevent you from stating the fact that substance Y melts at temperature Z.

Forms of protection fall broadly into two cases:

- Copyright for compilations
- A *sui generis* right for collections of data

As we have already emphasized, there are no general rules and the situation varies by jurisdiction. Thus we proceed country by country detailing which (if any) of these approaches is used in a particular jurisdiction.

Finally, we should point out that in the absence of any legal protection, many providers of (closed) databases are able to use a simple contract combined with legal provisions prohibiting violation of access-control mechanisms to achieve results similar to a formal IP right. For example, if X is a provider of a citation database, it can achieve any set of terms of conditions it wants simply by:

1. Requiring users to login with a password
2. Only providing a user with an account and password on the condition that the user agrees to the terms and conditions

You can read more about the jurisdiction by jurisdiction situation in the Guide to Open Data Licensing.

CHAPTER 2

Indices and tables

- *genindex*
- *search*

Index