Open Data Manual Documentation

Release 1.0.0

Open Knowledge Foundation

September 06, 2011

CONTENTS

This report discusses legal, social and technical aspects of open data. The manual 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 contributors or suggestions for additional sections and areas to examine.

CHAPTER

ONE

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 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 that makes it 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 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 positively impacts many different areas. In addition to the background, the manual also provides concrete information on how to produce open data.

1.1.1 Target Audience

This manual has a broad audience:

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

Most of the information provided currently 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 manual 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 manual.

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/

Copyright

1.2 Why Open Data?

Open data, especially *open government data*, are a tremendous resource that is largely untapped. Many different individuals and organisations collect a broad range of different types of data to be able to perform their tasks. Government is particularly significant in this regard, both because of the quantity and centrality of the data they collect, but also because of all that government data, most 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 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. The nature of innovation and new things is that it will come from unlikely places.

It is already possible to point to a large number of areas where open government data is creating value, and there are likely more. Some of these areas are:

- 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.

For transparency there are projects such as the Finnish tax tree and British Where Does my Money Go? that show how your tax money is being spend by government. Or there's the case 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 with all Danish public toilets so people she knew with bladder problems now trust themselves to go out more again. In the Netherlands, vervuilingsalarm.nl warns you with a message if the air quality in your vicinity is reaching 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 those parks. Services like mapumental in the UK and mapnificent in Germany allow you to find places you can live, given a certain commute time to your work place, prices of housing, 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 yearly in the EU alone. New products and companies are re-using open data. The Danish husetsweb.dk helps you find ways to improve the energy efficiency of your home, including financial planning and finding builders who can do the work. It is based on re-using catastral information, 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 translating algorithms, thus improving its quality of service.

Also for government itself open data is of value. Efficiency can be increased for instance. The Dutch Ministry of Education has published all of their education related data on-line for re-use. Since then the number of questions they receive has dropped reducing work load and costs, but the remaining questions now are easier to answer for civil servants as well, because it is clear where the relevant data to answer those questions 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 like the Wikimedia Foundation to execute their own tasks more effectively. This not only results in improvements of the quality of their data, but will also make the department smaller.

While there are numerous instances where open data is already creating both social and economical value in very diverse ways, at the same time we don't know yet what new things will become possible. New combinations of data can create new knowledge and insights, that lead to whole new fields of application. We have seen that in the past, such as when Dr. Snow in London discovered the relationship between drinking water pollution and cholera 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 meant a huge improvement in general health of the population. We will likely see that as well when unexpected insights flow from combining open data sets.

This untapped potential can be unleashed if we turn public government data into open data. Only, however, if it is really open, so if there are no restrictions (legal, financial or technological) to the 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, the public data needs to be open data.

1.3 What is Open Data?

1.3.1 What is Open?

This manual 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, reused 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, but to summarize the most important points:

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Reuse and Redistribution:** the data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, reuse 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 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, won't be open.

The key point to make 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 manual. It gives concrete, detailed advice on how data holders can open up data. Here 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 in 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 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 reusers of the data as early and as often as you can. Be they citizens, businesses, 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 reuse 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 the most important ones and (b) address them at as an early stage as possible.

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

1. **Choose your dataset(s).** Choose the dataset(s) you plan to make open, though keep in mind you can (and may need to) return to this step if you encounter problems at later stages.
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 catalogue to list your open datasets.

1.4.1 Choose Dataset(s)

Choosing the dataset(s) you plan to make open is the first step to take – though remember the whole opening up data process 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 what 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) - example see the 'What Data' section above.

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, are likely to have a good understanding of which data are valuable.

1. Prepare a short list of potential datasets that you would like feedback on. It is not essential that this list concurs with what your expectations are, the main intention is to get a feel for the demand. This could be based on other countries' *open data* catalogues.
2. Create a request for comment.
3. Publicise your request with a webpage. Make sure that it is possible to access the request on 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 decide 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, intellectual property rights in data 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 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 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 statue, 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 reuse that data.

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

- keep it simple,
- move fast, and
- 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 being online publication. There are many variations to this model. At its most basic, agencies make their data available via their websites and a central catalogue directs visitors to the appropriate source. However, there are alternatives.

When *connectivity* is limited or the size of the data are 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 teams 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 quite a bit of 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 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 of 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, as the demand increases, so does the supply. 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 at a time, 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 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 converting it into a new format, linking with other resources, data to be 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 OCTO's Data Catalogue, which allows data to be downloaded in CSV and XLS format, in addition to live feeds of the data.

via the data access protocol

DAP (Data Access Protocol) is a system for data transfer that was developed for use in meterology and climate science. The system was designed to enable third-parties to access sections of databases stored in some central location. Despite its origins in a particular field, the technology is very generic and can be adapted for data transfer in any area.

Implementing this technology can enable your agency to be experimental with its knowledge. For example, Australia's Bureau of Meteorology Research Centre provides the following disclaimer on its material:

> Please note that the following products ... do not currently form part of the Bureau's standard services in any way.

This example demonstrates that it is possible to provide data in raw form without incurring liability for others' use of that data.

via WebDAV

WebDAV, or Web-based Distributed Authoring and Versioning, is an attempt at making the internet a read/write medium. It is a widely supported open standard that supports locking and distributed authorship.

Providing a service such as this could be useful for situations where your agency would like to handle receiving improvements to data that it stores. The agency could provide its original data as the orginal source and then refer to higher-quality, but unverified derivative data source for users with different needs.

Offline methods

via optical media

Optical media, such as DVDs, are very cheap to produce. However, they tend to lack the capacity that would warrant the manual handling of distributing them. One exception to this is events. If you are hosting an event for developers, such as a hackfest or barcamp, optical media can be the best way to distribute a dataset for use in the venue.

via external hard disk drives

Hard disk drives can be very useful for data transfers in the terabyte range. To support this, you need to have some form of ability to receive funds to cover the purchase, handling and shipping of your data.

Be careful to make sure that you are not charging for the data. Instead, your fee should be as close to the actual cost of distribution as possible.

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 most discoverable.

The most prominent is CKAN.net. CKAN stands for the Comprehensive Knowledge Archive Network, and is a catalogue of all datasets in the world. The site makes it very easy for developers to find the material that they're seeking.

In addition, there are dozens of specialist catalogues for different sectors and places. Many scientific communities have created a catalogue 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 catalogue for the government's data. When establishing a catalogue, try to create some structure which many departments can keep their own information current easily.

Resist the urge to build the software to support the catalogue from scratch. There are many free and open source software solutions which have been adopted by many dozens of governments already. Investing in another platform will be a waste of resources.

There are a few things that most open data catalogues 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 catalogue as the region's catalogue, rather than the regional government's.
- Facilitating improvement of the data by allowing derivatives of datasets to be catalogued. 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 catalogue for hydrologists.
- Ensure that access is equitible. Do not create a priviledged level of access for officials or tenured researchers. This will cause resentment and ultimately undermine the goals that you are seeking to achive.

For civil society

Be willing to create a supplementary catalogue 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 embarrasment 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 catalogue 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 reuse.

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 you know may be interested in this data

Understanding your audience

Like all public communication, engaging with the data community needs to be targetted. 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 a 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 initative 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 a Creative Commons licence 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 sharable. 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 targetted 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 of driving signficiant 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 reusers
- Finding out more about demand for different datasets
- Finding out more about how people want to reuse your data
- Enabling prospective reusers 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. Following are several relevant examples.

1.5.3 Making things! Hackdays, prizes and prototypes

The structure of these competitions is that a number of datasets are released and programmers participate within a short time-frame, running from as little as 48 hours to a few weeks, to develop applications. 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 catalogue, 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 catalogue 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 catalogue. It included an online submission for applications, small but many prices instead of big but few prizes, 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 which 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 build on the reuse 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 licence and in a range of reusable formats. The 82 applications that 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 (CSO) 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 the social and economic advantages. CSOs that promote reuse 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 treating data such that it cannot be used for the identification of individuals.

Anonymization See *Anonymisation*.

API See *Application Programming Interface*.

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

AR See *Information Asset Register*.

Attribution Licence A licence that requires attributing the original source of the licensed material.

Attribution License See *Attribution Licence*.

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 the right of citizens to live without doubt of the consequences that might come if information of their private lives becomes public. The law protects the privacy (such as information about a person 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 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 that member states 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 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 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 licence 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*.

JSON A data exchange format that is often used on the web.

KML The Keyhole Markup Language, used for geospatial data exchange.

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 See opendefinition.org.

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.

PSB Public sector body.

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 Licence A licence that requires users of a work to provide the content under the same or similar conditions as the original.

Share-alike License See *Share-alike Licence*.

Tab-seperated values Tab-seperated 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.

XML A well understood text format for data and document exchange.

1.7 Appendices

1.7.1 File Formats

An Overview of File Formats

JSON

JSON is a very simple file format that is very easy for any programming language to read. Its simplicity means that 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 is built on, lets developers 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 for 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 could not determine how the remaining data must 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 reuse, 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 they come from. It should always be indicated, where you can find more information on these proprietary formats, 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 reuse 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,

reuse 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 most likely 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 affect 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, they are interested in.

There are some security concerns by 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 or even a single items (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 but 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 in 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,

> reuse 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, below we proceed country by country detailing which (if any) of these approaches is used in a particular jurisdiction.

Finally, we should point out that absent any legal protection many providers of (closed) databases are able to use simple contract combined with legal provisions prohibiting violation of access-control mechanisms to achieve similar results to a formal IP right. For example, if X is 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

(b) 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.

1.7.3 Making Personal Data Anonymous

Governments hold lots of personally identifiable and commercially sensitive information. This sensitive information necessarily restricts agencies' ability to share this information as open data. This article will go some way to introducing you to the option that you have to make it anonymous. Sometimes, just removing fields from a row is insufficient. It can be possible to use statistical techniques to identify individuals. This is especially the case when your data are combined with other sources of information.

Defining personal information

What counts?

This question is more complicated than what one would initially believe. Every jurisdiction has its own requirements and and every culture has its own norms about what counts as personal and private.

Things to look out for

Identifiers

Any number or other value that is used by a computer to identify individuals can be examined by an attacker. Typical identifiers include registration numbers, ID numbers, passport numbers, credit card numbers, IP addresses, and order numberr.

Very rare characteristics

Outliers are very easy to identify in a statistical manner.

Very specific descriptions

Specific descriptions make it very easy to identify individuals. As specificity increases, the number of individuals possessing a characteristic decreases. For example, there are fewer French than Europeans. While the specific information can be valuable, be wary of its effects in very small population segments.

Images

This can include photos of individuals, their property and other items of interest.

Biometric data

Biometric data are collected specifically to identify individuals. Therefore, it is highly likely that they will be sensitive. Biometric data includes fingerprints, retina, DNA, height, body markings. It may also include samples of handwriting.

Free text

Respondents' free text responses very can often identify who has spoken. Free text can be subject to word frequency analysis and other computational linguistic analysis.

Strategies for making data anonymous

Aggregate

When data are aggregated, they unable to be used to identify the sources of the data. That is, if we provide the mean household income for a street, we will not be able to identify the household income of any particular family. The downside of this approach is that aggregating data too far will impair analysts' ability to interpret the data.

Remove

A very simple approach to privacy protection. Here, we simply remove some field of interest that was originally collected. For example, we could omit gender, age, location or any other variable that has been collected.

Dither

Dithering results means to add a variation to every value within a sample, while attempting to maintain the integrity of the aggregate values. The goal is to prevent the the true value for any specific value to be deduced, but to enable statistical analysis to be carried out. For example, for geographic data, you could move points of interest to a random location within a given radius.

Top and Bottom Coding

Top and bottom coding means to replace extreme values of sensitive numerical variables with the weighted group mean for those values in order to mask outlying values which are potentially identifying. For example if the data is about the airline industry in New Zealand results from Air New Zealand might be replaced with the weighted group mean for the values in the group of data in order to ensure that the results from Air New Zealand could not be identified.

From the OECD:

> "It consists in setting top-codes or bottom-codes on quantitative variables. A top-code for a variable is an upper limit on all published values of that variable. Any value greater than this upper limit is replaced by the upper limit or is not published on the microdata file at all. Similarly, a bottom-code is a lower limit on all published values for a variable. Different limits may be used for different quantitative variables, or for different subpopulations."

Group

We can group multiple values together to protect individuals' privacy. Consider the following table, with the transformation appearing in the right-hand column.

132 cm	139.67 cm
143 cm	139.67 cm
144 cm	139.67 cm
144 cm	152 cm
153 cm	152 cm
159 cm	152 cm
161 cm	164 cm
167 cm	164 cm

There may be problems with this approach, as you will impact on the median and other percentile values.

Hash digests

Using a crypographic hash of a string can make it impossible for someone to determine what the original string was, while being able allow 3rd parties to check if strings they have are included. As they can apply the hash function to their own values, they can undertake comparisons without being able to access data that they don't already have. The transformation looks something like this:

researcher@example.org	c242dbe863aa0a38eacc72888fd41804
consumer@example.com	a99650df0d55169e0d9f1dc17194830f

References

- http://www.ihsn.org/home/index.php?q=tools/anonymization/techniques
- http://latanyasweeney.org/work/identifiability.html
- http://www2.stats.govt.nz/domino/external/omni/omni.nsf/23f076d733ded7e74c256570001d92b4/9476cd9e52a1d515cc2572
- http://stats.oecd.org/glossary/detail.asp?ID=7011

1.7.4 Barriers and Solutions to Open Data

This appendix was created from material generated from the SharePSI workshop of May 2010. The material is incorporated into the Open Data Manual with permission from Ton Zijlstra.

Introduction

This format does not take the narrative approach of the rest of the manual. Instead, bullet points taken mostly verbatim from the original meeting have been collected into themes. The themes include:

- Licensing
- Privacy
-

The responses in italics have been created by the Open Data Manual's authors.

Licensing

General sentiment

Licencing is a pain The fustrations over legalese can be minimised. There are now several publications which detail an appropriate license, including the *Open Definition*. Additionally, many governments have gone through the process of determining appropriate licenses. They will be able to share their experience.

Poor initial conditions

Absence of legal framework If a country faces the problem of not having a sufficient legal framework, then it can fall back on several non-legislative policy which has been developed. For Europeans, a large corpora of texts have been gathered by the ePSIplatform.

Unclear licensing This is less of an issue than it may have been during the meeting. The Open Definition provides an easily comprehendible guide to choosing an appropriate license.

Ignorance of licensing An ignorance of intellectual property, often compounded by an unwillingness to learn more, is a difficult problem to be solved. Most officials and politicians understand the need to be clear about what is acceptable, to protect all parties. You can use this general line of reasoning to talk about the specific case of data licensing.

Concerns around intellectual property Intellectual property rights can be unweildy. However, they are designed to give the owner of the rights control. Therefore, that control can be used to enforce open access with data that the government controls.

Licensing terms attached to datasets Complex licenses increase barriers to entry. Governments should seek to minimise the complexity to the contract terms that they impose. Preferably, those contracts should be *open* and easily understood.

Proliferation of semi-custom terms in licenses Open data is a relatively new field. This means that there are few established norms. As time progresses, we can expect to see a maturation in licensing schemes. Schemes will consolodate, reducing cost and complexity.

Incompatible open licenses See previous.

Share-alike licenses create silos Emerging community standards and peer pressure will reduce this problem over time.

Larger scale needed

Crossborder licensing needed for legal interoperability Streamlining legislation and policy between countries takes a long time. Progress is being made. However, even in cases where the legislation is not uniform, great value can be derived from doing something locally. Local businesses and community groups will be more than happy to use local data of their own area, irrespective of what is happening in other countries.

Concern to keep own national perspective on licensing is often bigger than actual international differences

This is an important consideration. It can sometimes be important to create a consistent local understanding of the issues before embarking on a process of international collaboration.

Most licensing initatives are single jurisdiction/sector This is at least partially to blame on intellectual property laws. They are all national in scope. Moreover, open government data are produced by governments. Government are also national in scope. With those two considerations in mind, it is perfectly natural that licensing is national. As time progresses, those policies will naturally streamline through a process of collaboration

Lack of harmonization between various EU level initiatives and projects, which risks fragmentation

While EU policy is never in perfect harmony, the trend tends is trending towards openness.

Privacy

Use of privacy concerns to prevent all discussion Privacy concerns can stifle brash action. This can be positive, as it can lead to greater consideration of the risks and benefits. With that in mind, all parties should reference their country's legislation to get a full picture of what is and is not restricted.

Privacy legislation Privacy legislation only covers private information. Much of government data is not private. Therefore, arguments should be made for extracting that data out of government, before moving to contentous issues.

Governments are concerned for people's privacy See above.

Concern around personal data See above.

Access

Digital divide Data require special analysis and interpretation before they are able to inform discussion. Only a small number of technical specialists can gain access intially. Notwithstanding that, those specialists are often in a position where they can make the data much more accessible than things currently stand.

Multiple languages required In regions where multiple and minority languages are spoken, efforts should be made to include the entire population. There is indeed a risk of segmenting the benefits to populations along linguistic boundaries. However, technology alo presents opportunities. Websites are much more easily able to be translated by machines than paper.

From access to new business

Increasing access to data

Access to data is political, not technical Political climate is changing. Many governments around the world are moving towards openness.

Getting data There are many more data catalogues available than was the case in 2010. Many of those catalogues provide easy access to the raw data in a direct manner. This shows that as the environment matures, the tools become more accessible to everyone.

Lack of standard open data policies This is changing, with governments using each others' policies to come together towards common standards. For example, New Zealand created its open access policy NZGOAL, which was then followed by Australia's AUSGOAL.

Too many data sources are not exposed yet As governments become more experienced with releasing data, they will be in a better position to release data that is more difficult to access. Lots of data are locked up in legacy systems. As those systems are replaced, local advocates are well positioned to make the case for open data to be considered.

Access to data is still largest issue Hopefully things have come some way since 2010. The quantity and quality of data releases have substancially increased in recent times.

Publishing data

Inconsistent and diverse formats Data formats are created to solve particular problems. Those problems are often sector-specific. Therefore, multiple formats are not a bad thing per se. Notwithstanding this, it is important to prevent this creep if at all possible. Data transferred over the web are moving towards the *JSON* data format. Where required, such as in the geospatial area, *KML*, an *XML* language is very popular. The most important factor is to provide the ability for third parties to easily read the data, which necessitates text rather than binary formats.

Transformation of data for publishing, ensuring correct tranformation When a government department needs to transform data in order for it to be used by the public, then there is always a risk of introducing corruption. Where possible, government departments should seek to provide access to the raw data. When it is not, automated processes should be created and followed for undertaking the transformation.

Lack of standards, or ad hoc standards Standards are emerging within communities of interest. For example, within the Linked Open Data cloud, there are requirements to create a full record at thedatahub.org with explicit and complete metadata.

Storing big data Many private sector providers have taken it upon themselves to solve this problem for governments. They often provide bandwidth and storage for public data at no cost to the data owner.

Finding and combining data

Lots of fragmented sources Public data are now being indexed by specialised search engines. This removes a large degree of the previous problems.

Lack of interoperability Interoperability concerns are particularly difficult when encountering non-open systems. Where possible, governments should seek to move to vendor-neutral, patent and licence free data formats.

Lack of info on what reusable data is there thedatahub.org is one of many services that provides information on public data. Its focus is on outlining exactly which data sets are released under open licenses.

Disparity of data sets The disparity of data reflects the disparities of the world. Some areas simply do not collect data that others do.

Unclear what data is there This uncertainty will hopefully reduce over time. There are now large volumes of open data available in several fields.

Limited quality of data Data can be of poor quality. Wherever possible, governments should seek to provide raw data. They can then work with third parties to build cleaner, more usable datasets for everyone.

Lack of findability of data See "Lots of fragmented sources".

No unifed data structures in Europe Data standards are often formed along sectorial lines, rather than national borders. If there is no prospect of consistency within Europe, try to build consistency within industries or disciplines.

Lack of metadata Many sectors are increasingly creating their own catalogues for their data. These catalogues often include excellent metadata. Where this has not yet happened, services such as thedatahub.org provide some ability to relieve these problems.

Reuse

Concern about usefulness of data Not all data are highly valuable. Yet, this fact should not be a general barrier the distribution of open data.

Unclear conditions for reuse Efforts such as the **:open:'Open Definition'** provide a measure of clarity within the fog. Unfortunately, there is a large proliferation of licenses used in the open data world. If you are considering to ignore a dataset because of licensing terms, make sure that you inform the owner of that. The owner may be in a position to amend their terms.

Limited user friendliess / information overload Data analysis is a technical skill. Yet, the skilled analysts are exactly the people who will be able to make data more user friendly and reduce the overload caused by floods of information. They are able to work with designers to generate lovely infographics. They are able to work with writers to consisely explain the trends and implications of data that are otherwise indecipherable.

Unclear data provenance Where the origin of data is genuinely unclear, provenance can be a significant concern. We need to know where data came from in order to be able to trust it. Without that trust, it is impossible to rely on it for analysis. However, there may be other uses which do not have such stringent requirements. Students could be given that dataset to practice their skills. The origins of the actual data are in this case irrelevant. All that matters here is that ther data are in a format that can be easily read.

Finding viable business models

Working with data is not easy This difficulty could be exactly where the business opportunities lie.

Starting local/small is not always possible, e.g. have to take MS at once for tenders Smaller businesses are also better placed to be nimble enough to take on less visible, riskier opportunities.

Scalability issues Concerns of scalability in open data business models are likely to be no worse than similiar concerns in other fields.

Data users still reluctant, mostly early innovators This is natural. As the open data movement matures, it will become more accessible to a wider audience.

Lack of business models The lack of business models currently is not in itself a reason to hold back on open data. Open government data can be used to lessen the costs of undertaking current business models, even without reference to any future effort that is yet unthought of.

Disruption of existing business

Current changing

Some public sector bodies have no choice but to charge for data Many agencies were created under a model of generating a revenue stream from the data that they collect. That fact in itself does not limit the applicability of the general argument surrounding pricing at marginal cost. Where the marginal cost of distributing data is negligible, the price should be zero.

Concern about reduced income to public sector bodies Income is likely to reduce if the current policy is to charge for access. Expenses may also decrease, as productivity gains from operating in an open manner are revealed.

Existing charging models See above.

What charging hinders

Unclear where decision on charging lies Responsibility for this depends on local circumstances.

Pricing models block market development by introducing arbitrary threshold for market entry One thing to note is that removing pricing may only lead to a small increase in activity in the near term. There remain very significant business risks for creating products from an open data market. Public sector agencies need to provide certainty that their open data stance will be long-lived.

Lower end of reuse market cannot exist for now The lower end of the market will take a fairly long time to develop, even when open data is widespread. The market participants at this segment have less capability and are unlikely to be able to execute new, profitable ideas.

Different perspectives

Some have stake in non-open data Conflicting interests are not unqiue to this area. Where interests do conflict, policy should seek to minimise any negative impact caused by this situation.

Media and journalism like to have exclusive access Providing open access to data does not provide open access to stories that emerge from that data. Data mining is complex and expensive. Data can be thought of as raw materials. Media outlets are positioned differently to refine these raw materials.

You cannot compete against free Yes, you can. Many businesses are built on providing a more convienent or more tailored service than a free alternative. Consider the case of bottled water.

Where not charging disrupts

Public sector bodies in direct competition with market with services based on their data which they also sell

Resellers will be nudged towards the value-added market segments. However, they also provide a convienent level of service and are also able to market their services effectively. Therefore, the disruption to the current data market may be radical, but is unlikely to be terminal.

Current markets seeing disruption (e.g. publishers) because of governments' publishing data sets with added value

See above.

Linked and Federated Data

Linked Data

- storage concerns
- search/browsing/exploration challenges
- manual revision challenges
- classification challenges
- extraction challenges
- interlinking of data challenges
- quality analysis challenges
- evolution/repair challenges

Jurisdictional

National authorities are neither financed nor mandated to create international interoperability

That may appear to be the case on paper, however in practice there are very strong incentives to undertake practices which lead to international interoperability. There is more collaboration between academics of the same discipline between continents than there is between academics of differing disciplnes at the same university. Industries are also highly globalised. There are often international standards, codes of practice and norms which lend themselves to consistency between countries. Lastly, national authorities are much more likely to adopted accept international best practice than take on the cost of developing their own standards.

General

No unified data structures

Linked Data is an exciting prospect. There is likely to be a large degree of reliance on this technology to be able to bridge current concerns.

Needed level/scale may superseed current stakeholders

Do not underestimate the need to be able to meet the needs of local stakeholders. Grand, beautifully designed policy frameworks are wonderful. Yet, to a family interested in the water quality of the river, a spreadsheet is much more practical.

Transition Process for Government

Lack of knowledge and awareness

General resistance to overcome

Cyncism is often accompanied with cost concerns and worries that job scope will increase without any recognition. Once the concerns are allayed, managed or resolved, then the resistance will be overcome.

Drivers are often external

Pressure from the outside can sometimes move governments the fastest. There must be external demand to justify that governments should supply.

Closed government culture

Government is heterogenous. While some aspects of government are very closed, others are not. Start by talking to the receptive listeners.

Barriers are often internal

The fact that a barrier stands in the way is not by itself a sound rationale for inaction. Instead, the costs and benefits of overcoming that hurdle should be weighed against the relative costs and benefits of other options. The relative position of different options should determine what action is undertaken, rather than the absolute value of any obstacles.

Lack of knowledge (data holders and users)

Local open data communities hold a wealth of knowledge. Officials should be able to trust their users. They should create relationships, just as commercial suppliers create relationships with their customers.

Lack of awareness

Awareness is quickly increasing. The open data movement is no longer new. This means that there is far more information to bring people up to speed than there once was.

Change is hard

Risk of overcomplicating issues

There is not a single type of open government data system. As the body of the Open Data Manual explains, it is possible to have a perfectly functioning open data that fits in well with many budgets, cultures and technical infrastructures.

Government is concerned by complexity

Governments only need to absorb as much complexity as they think is practical. If a huge range of policies need to be adapted to fit into an open data framework, then start with data sets which are simpler. If a legacy system would be too expensive to move into an open data environment, make that fact public.

Tensions exist between those sticking to old roles and those trying to adapt to new ones

Tensions between old and new are perennial. This fact in itself should not prevent any change from occurring.

Losing control, feeling disrupted

Government is concerned by losing control

> Government also has a mandate to act in the best interests of its citizens and residents. The fear of losing control is one based out of a lack of experience with this particular area. As more and more examples of useful things being created with open data, that fear of losing control will ease.

No inexpensive conflict resolution

The data owner retains ownership. For the public sector, it holds significant power if any conflict situation arises.

Data is power

Data is often unrealised power. Data is collected by governments for specific purposes. They do not have the flexibiity to experiment with using that data in ways which were not anticipated.

Data catalogues perceived as centralisation (loss of power and control)

Data catalogues are created simply to make things easier for consumers of the data. While this may be perceived as the loss of power or control, it is also the adoption of responsibility for support and upkeep by a central agency.

Government is concerned by a lack of security

See the security section, below.

Language

Different stakeholders speak languages, e.g. legal vs technical This is likely to be a transient issue. All parties will increasingly be able to communicate with each other as the open data movement matures.

Lack of common vocabulary See previous.

Seeking viable ways forward

Public service does not see own need for open data

Officials talk. As open government data spreads, news of its positive effects inside of government will be spread too.

Government's concern about open data's long term sustainability

The ability for the private sector to be able to consume, process and analyse large volumes of data will not decrease. Nor will its demand. The sustainability of individual open data intiatives is less certain. Public sector managers should seek to develop programmes which will be financially viable across changes of governments.

Uncertain economic impact

The economic uncertainties of open data are real. However, this justification for open government data is not purely financial.

Little empirical evidence

Empirical evidence is growing.

Security

Security threats If data owners are concern about security threats for distributing data openly, then they should third party services.

Fear of data manipulation Once data have been modified, it would be a misrepresentation for the data's modifier to claim that it is the original data. Therefore, if some harm is caused on the basis of that modification and/or misrepresentation, it's likely that the data owner would have some form of legal recourse to be able to insulate themselves.

Selective use of the data Effective communication is key. Data owners should be up-front with their data's limitations. This information can be included as a seperate file along with the source data or be displayed along side a download link or similiar.

Legal challenges Data owners are in a position to disclaim any responsibility for reliance on the data's accuracy in their terms of use.

Where does responsibility lie? Responsibility for security threats lies where it currently does.

Costs of transition

It's not as cheap as you may claim

There is more than one way to undertake an open data program. As we discuss within the manual, there are many alternatives to building a full service API. Many of those will be close to no cost.

Government procedures take a long time to change

They do. However, they are changing. Open data is no longer new.

No funds for transition

Start with changes which are likely to save money and increase efficiency. If there are data sets which different departments, or branches within departments need to go through a complicated process to access? If not, consider the difficulties that are currently required to access data from other levels of government. Each of these transaction costs impose a burden on officials.

The cost of transition falls with data owner, but revenue is gathered centrally by another agency

This is where a whole of government approach is required. There are circumstances where it is appropriate to look at a systems level to see the impact of current policy.

It's not as expensive as you fear We hope the Open Data Manual can go some way to minimising any costs whcih are incurred.

CHAPTER

TWO

INDICES AND TABLES

- *genindex*
- *modindex*
- *search*