

Ten Building Blocks of an Open Data Initiative

Successful open data initiatives involve more than just putting datasets online. This research note draws on an exploration of a number of different open data projects around the world to highlight key building blocks that open data initiatives commonly involve.

What is an open data initiative?

Open data initiatives work to make data more accessible and re-useable. There is no set template for an open data initiative, and they will vary according to their setting and scope. Initiatives can have a range of goals, from promoting transparency and accountability, to stimulating innovation and economic growth. Open data initiatives can be led by governments proactively releasing data they hold, or they may be led by private actors or civil society who collate relevant public data in ways that make it more accessible. Initiatives can focus on a specific topic (e.g. environmental statistics, transport data or international aid flows) collating data from across multiple jurisdictions, or initiatives might be centred on a particular geographical area or institution, seeking to increase access to all the datasets relating to a particular region, state or government body.

Building blocks

The following building blocks are not sequential. Open data initiatives may assemble them in different orders and different ways.

1. Leadership and bureaucratic support

In her exploration of open government data initiatives in the UK and US Hogge found that both *“a top level mandate”* from senior politicians, and an *“engaged and well-resourced ‘middle layer’ of skilled government bureaucrats”*, were essential to secure the release of open data (Hogge, 2010). Open data initiatives often need to overcome ingrained cultural barriers to sharing data, and to shift long-established practices of data handling. This benefits from strong senior leadership, and from skilled officers with the capacity and freedom to develop open data practices in agile, responsive and sustainable ways.

2. Datasets

Datasets are at the heart of any open data initiative. Open datasets need to be accessible (usually online), technically open, and legally open (Eaves, 2009). Many open data principles also highlight the importance of datasets being as authoritative, timely (published soon after collection), and as raw (granular data, shared prior to aggregation or analysis) as possible (Malmud & O’Reilly, 2007).

To be technically open a dataset should be in a non-proprietary format (e.g. CSV instead of Excel). It should be possible to load a well-structured open dataset into freely available software or online tools and to then sort, filter, search and manipulate it in a wide range of ways.

Initiatives might be able to export open datasets directly from existing internal databases and information systems. However, in many cases useful open datasets are the product of combining data from different systems and sources, and may need some manual preparation to remove personal information, or to make sure codes and field names used in the data are intelligible.

There are many different kinds of data (Davies, 2012a) and the availability of data to open often depends on the quality of digital data collection and good records management practices. Where an initiative is based around processing and republishing data, it is important for it to show the provenance of the data it is using and to link back to the original source of the data wherever possible.

Some open data initiatives select datasets to publish based on supply: working with 'low hanging fruit' and focussing on those datasets that are easiest to open first (Hogge, 2010), whilst others may adopt a demand-driven approach, surveying citizens (Both, 2012) or articulating use-cases to help decide what datasets to release first (Shkabatur, 2012).

3. Licences

A license sets out explicitly what someone who accesses a dataset can do with it. Datasets can be covered by a range of copyright and intellectual property laws. Without an explicit license, a user does not know if they have the legal permissions to share data further, to combine it with other data, or to build a commercial service off the back of a dataset.

Open data advocates emphasise the importance of licenses that have minimal restrictions, requiring at most attribution of the source of the data. They commonly oppose licenses that impose restrictions on commercial re-use of data (OKF - Open Knowledge Foundation, 2006). Incompatible licenses make it difficult to combine datasets, and so simple, permissive licenses are preferred.

An open data initiative may develop a licensing framework of its own, or might adopt one of the standard open database licenses that have been developed to apply to all its datasets (Hatcher & Waelde, 2007).

4. Data standards

A data standard describes the fields a particular dataset can contain, how they should be represented, and what conventions should be used for sharing dates, locations, categories and other common elements in a dataset. There are widely used standards for many kinds of data, from the General Transport Feed Specification (GTFS) for representing public transport timetables, to the International Aid Transparency Initiative (IATI) standard for representing data on aid flows and development projects.

When data is published using data standards it is easier to compare datasets from different providers, and to re-use existing tools with data. However, standards also limit what can be expressed in a dataset (for example, when their code lists do not

map the codes and categories used locally, and so any open data initiative has to make careful choices about which data standards it will use.

Standards also vary in their technical complexity. A CSV spreadsheet standard generates data which can be opened in spreadsheet software. More expressive standards like JSON and XML require specialist skills to use, but can support the creation of richer intermediary tools for accessing data.

Selecting or setting standards, and converting data into common standards is a key part of many open data initiatives.

5. Data portals

A data portal provides access to open datasets, hosting meta-data that describes them, and allowing visitors to search for relevant datasets.

Open data initiatives commonly provide their own data portals using specialist data catalogue software, although some may choose to curate a list of datasets in a common public data portal like TheDataHub.org.

Some data portals provide extra features, such as datastores and APIs¹, lists of interfaces and applications built with data, discussion and comment features, and in-built tools to visualise data. These should be offered in addition to access to download datasets, rather than instead of.

6. Interpretations, interfaces and applications

With open data third parties can provide their own interpretations or analysis in static reports and publications; they can build interfaces and visualisations of data to show trends and patterns in it; and they can create interactive applications that provide useful functionality.

In many cases, open data initiatives may provide or fund interpretations, interfaces and applications that make it easier to gain access to extracts from raw datasets, show what the data contains, or demonstrate the potential of the data. Showing visualisations and applications based on data can be important to secure political support for open data initiatives, and developing uses of their own data allows members of an initiative to identify opportunities to improve how it is published.

7. Outreach and engagement

Just putting data online is not enough to get it used. Outreach, community building and engagement is required. The 5 stars of open data engagement explain that an open data initiative should: be demand driven; put data in context; support conversations around data; build capacity, skills and networks; and lead to collaboration on data as a common resource (Davies, 2012b).

¹ API stands for Application Programming Interface. APIs provide a way for computers to query and fetch sub-sets of a dataset over the Internet programmatically.

Conferences, trainings, workshops, online Q&A sessions, 'hack days' and 'app competitions' are all methods that open data initiatives have adopted to support engagement. The US Data.gov programme also supports the development of thematic communities of practice around clusters of datasets. Hack days and app competitions, which challenge developers to come up with prototypes using datasets often produce striking results in a short period of time to show the potential of data, but only rarely lead to sustainable tools and services based on data.

8. Capacity building

ICT access, a supporting information infrastructure, digital literacy, and technical skills (Rahemtulla et al., 2011; Shkabatur, 2012) are all commonly required to make effective use of open data (Gurstein, 2011). Even for professionals such as statisticians or managers who work with data on a day-to-day basis, the tools and approaches commonly used in open data may be unfamiliar.

Capacity building often needs to take place on both supply and use sides of an open data initiative. Those managing data inside an organisation need to develop their capacity to provide well-structured, well-managed data, clearly documented with meta-data. Potential users of open data need to be made aware of the data that is available, and of different ways it can be used. Capacity building might also involve paying attention to skills for using data in research, campaigning, policy making or scrutinising government.

9. Feedback loops

Open data may attract many forms of feedback, from suggested corrections to the data itself, to annotations and additional contextual information, and substantive feedback on the issue the data describes (for example, feedback highlighting the spending reported in a dataset has not reached its intended recipients). Open data initiatives need to establish channels through which they can accept and work with feedback, either enhancing the data they hold, or taking action on the basis of feedback. Initiatives need to be clear to their users what they will do with feedback they receive.

10. Policy and legislative lock-in

Most current open data initiatives are voluntary. Governments or institutions involved in them could stop providing data at any time. Some initiatives have been exploring ways to develop a statutory footing by creating 'right to data' legislation, or writing open data clearly into contracts and policies.

Bibliography

- Both, W. (2012). Open Data - what the citizens really want. *The Journal of Community Informatics*. Retrieved from <http://communityjournal.net/index.php/ciej/article/view/814/918>
- Davies, T. (2012a). Untangling the data debate: definitions and implications. Oxford. Retrieved from <http://www.opendataimpacts.net/2012/03/untangling-the-open-data-debate-definitions-and-implications/>
- Davies, T. (2012b). Supporting open data use through active engagement. *Using Open Data: Policy modelling, citizen empowerment, data journalism* (pp. 1–5). Brussels: W3C.
- Eaves, D. (2009, November). Three Laws of Open Data (International Edition). Retrieved June 1, 2010, from <http://eaves.ca/2009/11/29/three-laws-of-open-data-international-edition/>
- Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday*, 16(2). Retrieved from <http://frodo.lib.uic.edu/ojsjournals/index.php/fm/article/view/3316/2764>
- Hatcher, J., & Waelde, C. (2007). Open Data Commons » Licenses. Retrieved June 1, 2010, from <http://www.opendatacommons.org/licenses/>
- Hogge, B. (2010). *Open Data Study. Transparency and Accountability Initiative*. Transparency and Accountability Initiative. Retrieved from http://www.transparency-initiative.org/wp-content/uploads/2011/05/open_data_study_final.pdf
- Malmud, C., & O'Reilly, T. (2007). 8 Principles of Open Government Data. Retrieved June 1, 2010, from http://resource.org/8_principles.html
- OKF - Open Knowledge Foundation. (2006). Open Knowledge Definition. Retrieved March 4, 2010, from <http://www.opendefinition.org/>
- Rahemtulla, H., Kaplan, J., Gigler, B.-S., Cluster, S., Kiess, J., & Brigham, C. (2011). *Open Data Kenya: Case study of the Underlying Drivers, Principle Objectives and Evolution of one of the first Open Data Initiatives in Africa*. Retrieved from <http://www.scribd.com/doc/75642393/Open-Data-Kenya-Long-Version>
- Shkabatur, J. (2012). Towards Open Government for Enhanced Social Accountability (How To Note).