An Open Analysis on Open Data

Anly Vathana, Independent Consultant

Dev Pramil Audsin, Senior Research Engineer, Orange Labs, UK

Introduction

Open data may be defined as some data which are available for reuse by everyone. Those data are not confined to patent; copyright etc. with the use of internet, the concept of open data has become a tool for businesses, analysis, discussion and decision making. The open data is booming as a next big thing around the globe. The open data concept is not new rather it can be considered as a replacing term for open definition which can be defined as a piece of data that can be used, reused and redistributed by anyone. In this paper, we have discussed about the sources of open data, issues regarding open data, formats for open data and presented some views and opinions.

Sources of open data

1. *Geo-data:* these data are helpful in formulating maps by considering the location of buildings and roads with respect to topography and boundaries.
2. *Cultural:* Data about cultural works and artefacts of the countries. These are data normally held by galleries, libraries, archives and museums.
3. *Science:* Data that is produced as part of scientific research in all categories from A to Z. The concept of open access to scientific data was institutionally established with the formation of the World Data Center system, in preparation for the International Geophysical Year of 1957-1958.
4. *Finance:* Data such as expenditure and revenue of government and information on financial markets such as stocks, shares, bonds etc.
5. *Statistics:* Data produced by statistical offices such as the census and key socioeconomic indicators.
6. *Weather:* Data obtained from satellite and other sources to predict the weather and climatic conditions.
7. *Environment:* Information related to the natural environment such as pollution, rivers, seas, mountains, volcanoes etc.
8. *Government:* Nowadays many national governments publish data catalogue for the sake of transparency of their plans and policies among the general public.

As mentioned above data come from many sources. In order to become useful for the end-user communities, raw data commonly go through various editing, aggregation and analytical stages. While researchers and academics may find the micro-data useful, policy and decision makers and the and the general public are more commonly interested in the easier to manage high-level aggregates.

Although open data has many sources, the major source are governments and multilateral global agencies such as World Bank, IMF, UNESCO, and the like who collect huge amounts of information for their various studies and future plans.

Many governments have voluntarily decided to release the data to the public because they believe wide and open access to such data will result in innovative applications being created and also new policies related to economic or other aspects of public interest can be developed. In United States the government open data has led to creation of innovative applications. United States was one of the first countries to opt for open data.

Necessity for open data

The need for open data are of manifold: public need to know the policies of government, data pertaining to health, environment will be of great help to the common person if shared using open data. Increased use of Internet increases the desire of people to access open data.

Issues of openness in relation to content and data

A user should be able discover the existence of the data. The expected users of particular data should in some form be informed of the existence of the data they are anticipating. The data must be available as a whole and cost effective in terms of reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form. The data should be easily accessible for research and analysis. It should be possible for the user to find the detailed information describing data ie., metadata and its production process. The data must be provided under terms that permit the reuse of data. And it should also allow the redistribution of data, including mixing with other datasets except some datasets pertaining to legal issues or similar one that should not be altered in any form. The users should also be able to access the data source s and collection instrument from which and with which the data was collected, complied and aggregated. Everyone must be able to use, reuse and redistribute. There should not be any discrimination against person or group. E.g. restrictions for use of certain purposes.

The user of the open data should have a provision to communicate with the agencies involved in the production, storage and the distribution of the data and should be able to share knowledge with other users.

Advantages of open data

For each category of data, the advantages may be specific and also pertaining to a particular community. For example, government data may improve public service delivery, corruption may be reduced due to transparency and better understanding of the plans, policies of the government. Likewise, the information regarding schools of a particular area can help parents to identify the right school for educating their children. This school information may not be useful for others.

In general, the advantages of open data are:

- creation of new socio-economic models
- better services
- lower IT cost

Open data project

As the open data has attracted many in the internet world, open data projects are under the go in many countries having been undertaken by various organisations. Open data projects mainly

focus on the presentation of the data to the end users in an easily accessible manner and sharable in a timely, secure and auditable way, visualisation of the data. Also some projects involve in the standardisation of data presentation format and tools so that data could be accessed from any environment.

The publishing of data involves certain cost. Some open data projects are funded and some not. In particular, government and science are often funded as projects. When the project ends, the funding stops. But maintenance and distribution of the data hasn't been budgeted for almost all the data sets we have today. So this incurs difficulty in maintenance of the existing datasets.

Many national governments have deployed data catalogues. Once the data catalogues are ready, they have to market it to the people. This is because the value of the data catalogue is realised only when it is being used by people.

When one begins an open data project, first they have to decide for what they going to build a data catalogue and why. They have to build based on the needs and demands of the end users. So before starting the project collect a statistics on who will be using the data, for what purpose and their demands. All successful open source projects build communities of supportive engaged developers who identify with the project and keep it productive and useful.

There are some critics such as that the existing procedures and datasets are not created, managed or distributed in an open fashion. There are no formal processes for managing and distributing updates. Building up a new sophisticated process incurs cost.

Since the publishing of data itself incurs cost, the data published should be the one that is of greater benefits to the end users. For example, the government data regarding new taxes levied will be much more important than the number of railway stations or such census. Also one has to make sure what benefit the publisher will get in addition to getting an ROI. In some cases even the invested amount will be obtained back. Its like the data is free for everyone. If the ROI figure seems to be sensible then those who publish data might be interested in funding new processes also. Also the organisations that wish to undertake the open data projects will come forward with better interest and involvement.

Some of the challenges regarding open data are:

- most require government data to be open, but not data are needed by everyone. Government data is free but when compared to cost of publishing, storage and maintenance, its not sensible to open up all government data that are not needed or neglected by most people
- Business people relying on open data needs guarantee regarding persistence, reliability and accuracy
- There need to be some benefit for publishers
- Data owners need to connect directly to reusers to understand their needs
- Analysis is not always easy
- Open data is not always good for decision making. Sometimes it leads to better decisions and certain times to poor decisions. For example, consider an open data catalogue regarding the performance of schools in a country, when a minor problem in a school is being exaggerated by a person who is one of source of information may lead the user to take poor decision of rating the school lower to other schools even though its teaching and other aspects are better than those of the other competing schools.

Opinions on open data

Obviously there are certain advantages and disadvantages in open data. But this highly depends on the type of data and its potential uses.

- Government uses the public money (money collected in the form of taxes etc.) for work, so those data regarding how the money is used need to be transparent to the public.
- Something that have happened (facts) cannot be copyrighted.
- Data are an important enabler of socio-economic development (health care, education, economic productivity, etc.).
- In scientific research, the rate of discovery is highly influenced by the better access to data. Also sponsors of research do not get full value unless the resulting data are freely available.
- Data on genomes, data on organisms, medical science, data on environment are important for human in some perspectives of day to day life.
- Restrictions on data re-use create a situation where potential users can exclude one another.
- Government and private enterprises should co-operate to optimally serve open data to people.

Open data standards & Modern technology - based on the Internet - has given rise not only to many useful automated tools for the collection, transmission, and processing of information, but also placed great emphasis on the use of open standards to facilitate exchange of data and metadata.

Data comes in all shapes and sizes. The “open” in “open data” is predominantly about the licensing terms that are applied to data, and hence how it can be used by others.

Open standards are standards that are developed through a fair, transparent, collaborative process, available under a royalty-free license. Open standards may apply to data formats, to the protocols and APIs that are used to pass around information, and to tool configuration.

To be most useful, open data should be made available using a format defined in an open standard, for example as an XML, JSON or RDF format, and should be delivered over a protocol defined in an open standard, such as HTTP, as well as being licensed with an open licence.

1. Available on the web (whatever format), but with an open licence
2. available as machine-readable structured data (e.g. Excel instead of image scan of a table)
3. use non-proprietary format (e.g. CSV and XML)
4. use open standards from the World Wide Web Consortium (W3C) such as RDF and SPARQL to identify things, so that people can point at your stuff
5. link your data to other people’s data to provide context.”