FACT SHEET - Creating reusable machine-readable data

This fact sheet provides guidance to agencies on releasing data in formats that are machine-readable and allow for easy reuse under the WA Whole of Government Open Data Policy.

The principles of open and machine-readable data

Data published in a manner consistent with the principles of open data shall be machine-readable, licensed under a permissive license (eg Creative Commons) and available in easily accessible and 'open' data formats. This will minimise the obstacles to reuse.

As the name suggests, *machine-readable* data is data which can be read and interpreted by a computer program without the need for manual human intervention. The data is structured in a simple and consistent open data format that permits easy interrogation by computer code and does not require the purchase of a specific piece of software or operating system in order to access.

Currently most government data is released in formats that are tailored for human consumption and re-use (eg tables or graphs in PDF or Word documents or within HTML web pages). For this data to be reused by others, it usually needs to be laboriously re-entered by hand, or copied and manipulated into a useable format, often adding significant time and effort to reusing the information.

Releasing this same data in an open and machine-readable manner makes the data much quicker and easier to reuse and can be as simple as exporting the raw data from the source in an appropriate format (such as Comma Separated Value [CSV]) or spending a few minutes pasting the data into a spreadsheet application, cleaning the data to remove unnecessary formatting, and saving it back out in an open format.

If you are publishing data in a non-machine-readable open format (eg a written report or publication, or using a non-open format) it is recommended to also provide the raw data alongside it in open formats.

Other resources: the W3C's first working draft of the Open Data Standards.

What are the appropriate formats for my data?

For tabular data, such as financial reports, statistics, the preferred format is CSV (JSON and XML may also be acceptable). Avoid the use of the proprietary .XLS(X) formats that Excel creates by default. Although if your agency already has a large collection of data already formatted as .XLS(X) then, at least initially, providing that is preferred over not providing any data at all.

For textual data, such as reports and publications, try to use HTML, plain text (.TXT), or accessible PDFs, rather than Word documents (.DOC[X]). Where structured or tabular data exists within the textual data it should be published separately in CSV or another appropriate format.

For data that has a geographic or location element to it GeoJSON, CSV, KML, or GML are all good open formats for publishing. Shapefiles and TAB files should be avoided where possible due to the overly complex, proprietary, and aging nature of the formats.

Other resources: Open Data Handbook – File Formats; UK Government Service Design Manual – Choosing appropriate formats.

Guidance on publishing CSV files

CSV files are easy to produce and are the preferred format for publishing machine-readable tabular data. CSV files are a very basic "flat" file format and can be produced using software already available to Government employees, such as Excel, OpenOffice, and Google Docs.

Follow these three simple rules to create clean, machine-readable CSV files:

1. Remove any comments, metadata, and aggregate statistics intended to make the information easier for people to read.
2. The first row of the file should contain column names describing the data contained in each column.
3. Data should be free of formatting and any special characters (eg numeric or currency cells should just be numbers, without any formatting or non-numeric characters). This includes the use of colour, bolded and italicised text, and other visual aids.

These rules are necessary because formatting is not carried forward into CSV files. In addition, the inclusion of extra information (eg metadata and human-readable comments) makes it harder for computer code to parse and read the data through analysis and programming techniques.

You can read more about CSV files on Wikipedia's guide to the basic rules and examples of CSV files.

Checking the quality of your data

Before uploading your data to data.wa.gov.au it is good practice to run a final check over it to validate that it contains good quality machine-readable data (eg numeric columns only contain numbers, the date formatting is consistent, etc).

Fortunately there are web-based tools such as the OKFN Labs Good Tables service that automate much of this process. Good Tables allows you to upload (or link to) a CSV/Excel formatted file and have it run a set of automated checks across the data to ensure that it meets some of the common standards for machine-readable data.

Other resources: The Open Knowledge Foundation - CSV files explained; Clean sheet: Releasing data or statistics in spreadsheets; School of data: A Gentle Introduction to Data Cleaning.

The 5-Star Open Data Model

The World Wide Web Consortium (W3C) the standards body for the internet, has developed a five star model to describe different characteristics of open data and its usefulness for people wishing to reuse it. It is being used globally as a model for assessing data readiness for reuse.

As outlined above, in Western Australia most government data currently made available to the public is done at a one or two star level. The three star level is considered the minimum standard for release of the government's public data for reuse. This requires some work surrounding data formats as outlined above such as non-proprietary and machine-readable formats. The three star level also involves making the data accessible via the data.wa.gov.au portal, and licensed in a way that promotes dissemination and reuse of the data (eg one of the licences under the Creative Commons licensing framework).

5-Star Deployment Scheme for Open Data

Step 1: Put your data on the web with an open licence
An open licence means that people can easily understand the terms under which your data is available for re-use.

Step 2: Make it available as structured data
For example, an Excel spreadsheet is more useable than a scan of a table in a PDF/DOC, and saves users from manually entering your data into their spreadsheet.

Step 3: Use open, standard formats
Non-proprietary formats can be accessed by any software – for example, you can save an Excel file as a CSV file, which is an open format.

Step 4: Use URIs to identify data
Uniform Resource Identifiers (URIs) are a type of web link. They make it easier for your users to point at and draw upon on your data.

Step 5: Link your data to other people's data
Linked data uses a common structure and format, so your data is standardised, and can more easily be joined with other datasets.

Excerpt from ict.gov.nz's Guide to the 5 Star Open Data Model licensed under CC BY 3.0 N