Open Data Release Checklist

Expanded detail about each item on this list follows the list below. The idea is this resource could be used as a checklist to ensure all that should be considered before release, has been. But if you want further explanation click on the "read more…" link.

Identification

- ❑ Data Inventory (know your data assets) (read more…)
- ❑ Examine website for information with data behind it (read more…)
- ❑ Review OIAs and data requests (read more…)
- ❑ Engage and ask data user communities (read more…)

Review

- ❑ Prioritise data for release (read more…)
- ❑ Assess ownership and rights over data (read more…)
- ❑ Indigenous data sovereignty (read more…)
- ❑ Privacy Impact Assessment (read more…)
- ❑ Political sensitivity review (read more…)
- ❑ General risk assessment (read more…)
- ❑ Review standards applicable (read more…)

Authorisation

- ❑ Gain appropriate approval to release (read more…)

Preparation

- ❑ Data cleaning (read more…)
- ❑ Confidentialisation (read more…)
- ❑ Metadata (read more…)
- ❑ Data dictionary and Entity Relationship Diagram (read more…)
- ❑ Documentation including:
  - ❑ Context about collection and purpose of the dataset (read more…)
  - ❑ Explanation of the data (read more…)
  - ❑ How to use the API (if API available) (read more…)
  - ❑ Query examples (if released through an API) (read more…)
- ❑ Determine release frequency and process (read more…)
- ❑ Maintenance planned (read more…)
- ❑ User engagement and testing pre-release (read more…)

Implementation

- ❑ Open Data formats (read more…)
- ❑ Open API standard (read more…)
- ❑ Licensed for reuse (read more…)
- ❑ Consider need for additional and prominent disclaimers/warnings (read more…)
- ❑ Data published (read more…)
- ❑ Documentation published (read more…)
- ❑ Feedback option available (read more…)
- ❑ Invitation for reuse stories (read more…)
- ❑ Listed on a central catalogue/portal (read more…)
- ❑ Availability promoted (read more…)

Expanded sections:

- ❑ **Data Inventory (know your data assets)** (Top)

Good quality open data will come from well managed data, data which is managed as a valuable asset. In order to manage data well, you must first know and understand what data you have. Therefore, creating and maintaining a data inventory (catalogue) is an important step. It might seem a simple step, but it can be a bigger job than expected.

Here are some thoughts to guide your data inventory development:

- Focus on one business unit at a time
- keep it simple, don't create too many columns (this can lead to an overwhelming level of information gathering about each dataset and slow progress) - more on this later…
- Aim to make it public so that people can discover and understand what data is held by your agency. This means describing the datasets in a way that can be understood by both the public and the business unit.

Regarding columns in the inventory - **in New Zealand** we are developing inventories that include the following:

- DCAT standard metadata columns - used by the CKAN catalogue at Data.govt.nz
- **plus** columns for: Open? (Y/N); Could be open(Y/N); What's required to address before release? Estimated date of release
- **plus** a few columns of choice by each agency that met their needs to manage the data in their context.

View an example data inventory file [Link].

The data inventories are created as machine-readable CSV files. This enables an easy way to semi-automate the update of Data.govt.nz through simple script to select rows where Open?=Y and convert them into a "Data.JSON" file saved on the agency website, where Data.govt.nz can harvest your listing to keep it up to date. Therefore by updating the inventory you can also update your open data listings on Data.govt.nz.

Other benefits

The process of carrying out a data inventory can also lead to a number of other benefits, such as:

- surfacing of duplicate datasets (or multiple different versions of the same thing)
- a business unit realising that another business unit has data they want but didn't know existed
- discovery of ownership issues with data sourced elsewhere - that need to be sorted out
- more conversation about whether data could be made open
- more awareness of data, let alone open data, across the organisation
- conversations about the benefits of more transparency
- evidence for more resources to manage the ever growing volume of data

- ❑ **Examine website for information with data behind it** (Top)

This can be part of your data inventory process but worth doing as a separate exercise. Examine your website content for information that must have data behind it, such as tables and graphs. These may be on pages or as part of published reports for download.

This exercise helps you realise where data might be sitting in your organisation that people have forgotten about. If the information has been browsed or downloaded then you have evidence of interest in the data behind it.

- ❑ **Review OIAs and data requests** (Top)

Official requests for information or data can be a clear indication of demand that could be proactively met through open data. This source of information could also help you quantify savings that could be made by releasing the data, by reducing the time spent on responding to requests.

Some agencies now have the practice where if an official request leads them to prepare and release data to someone, they then publish the data for all to access going forward.

- ❑ **Engage and ask data user communities** (Top)

A great way to discover what data your agency has that is of value to others is to talk to people. Get involved in conferences, and not just data related conferences, but also conferences in the charity sector for example, to learn what others are doing and to think about what data may help from your organisation.

Also get involved in any hackathons like GovHack (join a team!), open society events and Open Government Partnership gatherings.

Either initiate or join in Meetups and find out projects people are working on. There are four open data meetups around New Zealand, Auckland, Wellington, Christchurch and Dunedin.

Wherever possible, keep engagement informal and around food. Break down the perception of "us and them" (government, non-government) and meet just as fellow humans passionate about putting data to use and learn from each other.

Be bold and meet civil society where they are and join the conversation with an open mind, for example New Zealand's Open Government Ninja forum where action-inspiring conversations have been going for over a decade.

Indigenous data sovereignty

Indigenous peoples' rights to data are set out in the UN Declaration on the Rights of Indigenous Peoples (UNDRIP). Article 18 stipulates that Indigenous peoples have the right to participate in decision-making in matters that would affect their rights, in accordance with their own procedures.

That participation should be from designing the collection through to the publication and use of data, with the opportunity to work towards beneficial outcomes for indigenous peoples. Partnerships between data publishers and indigenous peoples help address all information needs and manage expectations and responsibilities between partners.

When using data, especially when making decisions affecting indigenous peoples, we have an obligation to ensure they have had appropriate participation in the collection and care of the data.

There may also be obligations under local treaties or law to consider (for example, the Treaty of Waitangi in New Zealand).

- ❑ **Prioritise data for release** (Top)

Of course, all data that's non-personal and unclassified should be proactively released, but it can't all be magically released at once. It does take time and effort, and there are a number of ways to prioritise which data to work on releasing first, it just depends on where your organisation is at on its open data journey.

If you're just starting out then finding a dataset in high demand and that is likely to get used straight away to make an impact will be a good candidate to focus on first. Once the data has been used, tell the story. When people see how open data can be used for the benefit of others the resistance to open data starts to wane.

Another approach can be to use the expectation to openly release data as a lever to get data into better shape - for your own organisation's purposes. Therefore choose a valuable dataset that needs lots of work to improve its management and quality - and use the fact you will be releasing it publicly as a driver to get better governance and process wrapped around it.

- ❑ **Assess ownership and rights over data** (Top)

It's important to establish the copyright over the data. Is it yours to release? Or is it actually owned by someone else? Has your organisation acquired the data under licence from someone else?

A tool to help you through what you can and can't do if the data is not fully under your copyright is the Review and Release flowchart in the New Zealand Government Open Access and Licensing Framework (NZGOAL).

In order to release data you need to be sure you own it, and even then there may be reasons why it can't be released. Therefore you need to review:

- how the data has been acquired?
- were there any caveats on the data being provided (for example, confidentiality)?
- was the data purchased? If so, are there any restrictions to releasing as open?
- Was the data sourced from a third party organisation with restrictions on its use?
- Is the dataset a blend of your data and somebody else's? Are there restrictions on the use of the data from the other party?

If the data, or part of it is actually owned by someone else it does not necessarily mean it can't be released, you can try communicating your intentions and get their agreement to do so.

- ❑ **Indigenous Data Sovereignty**

Has the data been considered in the light of the 2007 United Nations Declaration on the Rights of Indigenous Peoples? Is there local treaties to consider? for example in New Zealand, Te Tiriti o Waitangi.

Your review should consider how indigenous communities are represented in the data; what input have they had into the design of the data collection? What is the likely impact on them in publishing the data? Should there be further consultation before release?

Ideally a relationship should be established at the beginning of the data lifecycle, the design of the data collection, so that Indigenous people can be empowered to benefit from the data for their desired outcomes. By the time it comes to potential publishing of the data a collaborative approach to release will have already been worked out.

- ❑ **Privacy Impact Assessment (if required)** (Top)

Always consider whether you need to do a privacy impact assessment. You may not need to if the data is something like the location of road signs, but it's a good habit to always think about whether there might be privacy concerns.

In New Zealand we have a very useful Privacy Impact Assessment Toolkit provided by the Office of the Privacy Commissioner, and it is recommended to use this to determine the level of impact and the appropriate measures to safeguard privacy.

A privacy impact assessment may not be required but you should always think through whether or not one is necessary.

You can later come back to these people you meet to ask for feedback on your beta API or trial dataset you're preparing for release.

- ❑ **Political sensitivity review** (Top)

Its difficult to give examples, but you just need to be aware of what's going on in the world outside. Sometimes it just might not be the best time to release your data when there are tensions in the public domain on a related issue. You may look to delay release of the data, or manage the release through extra effort to proactively communicate about it.

A political sensitivity review may not be required but you should always think through whether or n

ot one is necessary.

- ❑ **General risk assessment** (Top)

A general risk assessment covers any issues that don't come under privacy or political assessments, such as commercial sensitivity, potential security concerns or the risk of the data being used to cause harm. The issues identified may not mean the data shouldn't be published, but there may be actions required to manage the impact.

A general risk assessment may not be required but you should always think through whether or not one is necessary.

- ❑ **Review standards applicable** (Top)

Wherever possible you should adhere to open standards when releasing data, this is to ensure maximum access to the data that does not rely on any specific proprietary software.

As you'll see after you have worked through this section, the sooner you consider standards the better because the impact downstream in the data lifecycle can be significant and may mean your data is less valuable to users if it is not consistent with other data they wish to combine it with.

What needs to be standard?

Data formats

The data should be published in an open, machine-readable format, and if possible, multiple formats is preferable (again to increase accessibility and usefulness of the data). Examples of open data formats are...

Tabular data:

- CSV

- XML
- JSON

Geospatial data:

- GeoJSON
- KML
- GML
- Esri Shapefile

A CSV file may also serve as a geospatial file if it contains two columns for longitude and latitude values.

To learn more about open geospatial formats, go to GIS Geography.

Metadata standards

When releasing data, you should check to see if there is already a standard way to describe your data through metadata. The benefit of doing this is to ensure that your data is consistently discovered along with other like datasets through searches; to help data users to understand whether your data is likely to be compatible with other data they are looking to combine it with.

Content standards

There are two levels of content standards, record level and variable (attribute) level.

At **record level** we are concerned about what pieces of information we collect to describe the event/instance that the data is about. For example, datasets on tsunami evacuation zones may want to contain a minimum of evacuation zone type; location; height; shape area; and shape length.

So check to see how others are releasing the data, and where appropriate, talk to other agencies in the same industry.

Variable level content standards are concerned with having consistent and comparable values in variables within a record such as name, address, date of birth, gender etc..

For example in statistical data about people, when comparing based on gender, it's important that all data has been collected with the same list of options and have values are consistent like having “Male” instead of just “M”.

Data Collection Standards

For data to be comparable it’s important it to be collected consistently, not just in terms of content standards, but also the method of data collection.

For example, water quality. What are the types of particles that are typically collected by sensors, and what is the typical level of accuracy of the sensors? What other factors are considered and recorded to determine water quality such as ecological health?

Good context with data helps others to understand how data has been collected and can enable a sector to move towards their own standard if no useful international standard already exists. An example of helpful context on how data is collected is this description by the Waikato Regional Council on how they measure water quality at https://www.waikatoregion.govt.nz/environment/natural-resources/water/rivers/healthyrivers/how-we-measure-quality/

API standards

It's important that your data API adheres to open standards. If not, your data may be inaccessible to people that do not have/cannot afford the proprietary software required to connect to your API. A data user should be able to query your API using commonly known commands and query language.

For more information about API standards, go to:

Digital.govt.nz
https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/application-programming-interfaces-apis/

- ❑ **Gain appropriate approval to release** (Top)

Every agency may have a different authorisation process, and different points along the journey to gain approval to release data. In some cases you may need to get approval very early in the piece when working through this checklist. In other cases, a governance committee may expect to know the outcomes of privacy, political and general risk assessments and what is proposed to manage any risks. Therefore the assessments should be done before seeking authorisation.

One possibility is to work towards developing a procedure that is approved, and where if certain criteria are met in assessments the release is "pre-approved", with only exceptions needing to be formally approved.

- ❑ **Data cleaning** (Top)

Data cleaning is the process of correcting obvious errors or making things more consistent (eg. putting 'city' in the City variable/column rather than in Address Line 2).
The amount of cleaning depends on the size of the data and how it has been collected. Ideally cleaning would be done in its primary location (eg. in the database behind a business/service delivery system), rather than after you have extracted data out separately for release, otherwise you will only have to repeat the process when the data ise re-extracted to release an update.

- ❑ **Confidentialisation (Anonymisation)** (Top)

Organisations have different requirements for protecting data confidentiality, privacy, and security so people, households, and organisations can't be identified without their permission.

Protecting privacy is critical to maintaining trust and confidence in open data, as well as keeping people safe from harm. There are different statistical methods that can be applied to data to reach a confidential state, to learn more go to:
https://www.data.govt.nz/manage-data/privacy-and-security/understanding-data-confidentiality/.

- ❑ **Metadata** (Top)

Metadata is a small amount of data that describes your dataset. Think of it as a header page on a batch of documents. It helps people browsing the data to assess whether it is likely to be what they are looking for. It should include information like (but not limited to):

- Title of dataset
- Description
- Creator/author/organisation
- Date created
- Date last modified
- Version
- Licence
- Data format

The description is key to people understanding what the data is about, so a concise description, free of jargon is important.

How much else you include in the metadata depends on whether your data is accessible by API or by browse and download. If the data is accessible by browsing and download, then you can also make available, either on that page or by linking additional documents, contextual information about the collection and purpose of the data and any other information that will assist a good understanding of the data and supporting effective reuse of it.

If your data is accessible via API, then you may want to add this contextual information in as metadata. When someone explores the data API "endpoint" with a view to use the data, they will see the metadata, so they can learn about the dataset there before the begin querying the data or programming an interface to it in their software application.

- ❑ **Data Dictionary and Entity Relationship Diagram** (Top)

Data Dictionary

A Data Dictionary is a technical document describing each attribute (or variable) in the dataset. It typically shows:

- what kind of value it holds (date, number, alphanumeric text, binary (Yes or No) etc.)
- any limit to the number of characters entered
- whether its mandatory to enter a value
- whether it is a "primary key" (unique identifier for that record in the database
- whether its a "foreign key" (the value needs to equate to a primary key in another table).

The foreign key is only necessary where multiple related tables are made available.

A data dictionary is important information for a software developer to make good use of the data.

When data is released as a CSV file format, having been exported from a database, some of these details are not necessary but can still be helpful to understanding the data.

Examples of Data Dictionaries:

Building outlines data (New Zealand) at
https://nz-imagery-surveys.readthedocs.io/en/latest/published_data.html

Alternative disputes resolution data (Ontario, Canada) at
https://www.ontario.ca/data/alternative-dispute-resolution
(scroll down on the right, below "Data description")

Entity Relationship Diagram

When multiple related tables of data are made available an Entity Relationship Diagram can help data users understand the structure of the data and the relationship between the tables more clearly. It will help them write data queries across the tables that will return results that make sense.

Examples:

New Zealand Charities Register, at their information page for software developers, have a link to an Entity Relationship Diagram, along with a Data Dictionary, what you need to know about using the web service (API), the formats of the data available and example queries from different software tools.

- ❑ **Context about collection and purpose of the dataset** (Top)

To help data users understand the data and to use it appropriately, it's important to provide good context information, such as:

- the purpose the data was originally collected for
- how it was collected
- important concepts that might need explaining for clarity
- any limitations (such as the accuracy level of sensors or known bias in samples)

And anything else that you can think of that will help data users make the best use of the data (and manage any inappropriate use through misunderstanding).

Sometimes all you need a paragraph, this example at the source of biosecurity plant pest data at Marlborough District Council conveys a lot of useful information very succinctly:

> "This Council dataset records where plant pests have been found in the past by Council staff or through confirmed reports. Strong caution is advised as there may be other infestations in Marlborough that have not yet been captured in this dataset. If you suspect any other infestations that are not marked on this map, please contact the Biosecurity team at Council. This dataset does not distinguish the density or nature of the infestation at a given site. A marked area could range from heavily infested to only scattered plants. Either way, there is still a spread risk from both. Data updated weekly."

In just one paragraph they have communicated how the data is collected; cautioned about a weakness that there might be gaps; how you can help them fill those gaps (both managing the risk and inviting engagement and contribution from data users); explained how to interpret what you will see in the data; and how frequently the data is updated.

More examples:

Helpful context on how data is collected is this description by the Waikato Regional Council on how they measure water quality:
https://www.waikatoregion.govt.nz/environment/natural-resources/water/rivers/healthyrivers/how-we-measure-quality/

Building outline data from Land Information New Zealand (LINZ):
https://nz-buildings.readthedocs.io/en/latest/introduction.html?fbclid=IwAR1eX2_ozq5TamgkQsyaUMpit1uys3DeADdf_zikqrOG86CDkH7iGc9Gm58

- ❑ **Explanation of the data** (Top)

In addition to context, it may be helpful to describe what some variables mean, especially in scientific data. It may be possible to have one reference piece that supports the understanding of a number of related datasets.

- ❑ **How to use the API (if API available)** (Top)

APIs can be subtly different to interact with, having slightly different commands and syntax. Having sum guidance for developers on how to use the API can speed up their access to the data considerably, saving their time and achieving impact more quickly.

An example of the kind of help needed to help developers going is at the New Zealand Charities Register open data:

https://www.charities.govt.nz/charities-in-new-zealand/the-charities-register/open-data/

- ❑ **Query examples (if released through an API)** (Top)

Building on the idea of providing guidance to use your API, it is very helpful to data users, especially those unfamiliar with your specific API platform, to provide some examples of how to query your API. This will assist them in accessing the data and getting on with using it much more quickly, and for some, examples may be all they need.

The New Zealand Charities Register API user documentation does this well:

https://www.charities.govt.nz/charities-in-new-zealand/the-charities-register/open-data/

- ❑ **Determine release frequency and process** (Top)

When considering the frequency of updating the data you need to consider what work and/or process is required to update the released data. You also need to consider what is useful to the consumers of the data. This may require some engagement with data users early on so you can prepare to meet expectations that lead to good use of the data.

Some different examples to consider:

- **Statistical data** - consider the level of work required to prepare for release, and determine whether there are existing release schedules and data preparation processes that the dataset would fit in with
- **Data from sensors** - is the data most valuable in real-time? If so, can it be released in real-time? Or does the situation call for some data cleaning or processing before release?
- **Counts** (eg. traffic counts) - what frequency is most useful for data consumers? hourly, daily, weekly or monthly?
- **Registers** - how often do changes occur? How often is the register accessed? Are daily updates required by users or would weekly or monthly be sufficient?

- ❑ **Maintenance planned** (Top)

As you prepare to release the data, ensure that the appropriate process and procedures are in place to ensure the regular update of the data. In some cases this may be closely tied with a business process or it may mean setting up an automated update process. In other cases, you may need to ensure a way of alerting someone that an update is expected.

For some data you may need to consider how you want to present time-series. If publishing static files, do you want to publish snapshots in time in separate files or combine additional time periods into one file?

- ❑ **User engagement and testing pre-release** (Top)

One way to get impact from releasing open data is gather a community around the data along the journey towards releasing it. In addition to helping you design a good data release, this community can also help you test your data and the access to it (eg. API) and give you feedback from a user perspective before you go live.

- ❑ **Open Data formats** (Top)

An open data format is a format that is format for storing digital data, defined by a published specification usually maintained by a standards organization, and which can be used and implemented by anyone. An open format does not require specific proprietary software to access the data, and is machine-readable.

To read more about open data formats and a table of examples, go to:
https://www.data.govt.nz/manage-data/policies/nzgoal/guidance-note-2/.

- ❑ **Open API standard** (Top)

An open API is one that conforms to requirements defined by a published specification usually maintained by a standards organization. An open API does not require any proprietary system to access it, and uses well known commands and standard protocols for interaction.

To read more on open API standards and guidelines go to:
https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/application-programming-interfaces-apis/

For information on the qualities of a good open data API from the user's perspective, go to:
https://www.data.govt.nz/manage-data/releasing-data-on-data-govt-nz/qualities-of-a-good-open-data-api/

- ❑ **Licensed for reuse** (Top)

With the notable exception of the US Federal government, publications by governments around the world are subject to copyright by default. This means when publishing data on the web, by default the ability for people to reuse it is heavily restricted.

So to make data open, it is critical to license it for reuse, to explicitly give permission that overrides copyright restrictions. Some countries have written their own open licences (such as the UK Open Licence), but many countries have a policy of using the Creative Commons Licences, and provide guidance on how to apply them.

For an example of policy and guidance on how to apply Creative Commons Licences to government data, information and publications, refer to the New Zealand Government Open Access and Licensing framework (NZGOAL).

- ❑ **Consider need for additional and prominent disclaimers/warnings** (Top)

At the point of release to reduce reliance risk to people if their safety is at stake, you may need to consider the need for additional disclaimers or warnings (e.g., 'this data should not be used for X, Y, Z purposes without independent validation; depends entirely on the context as to whether any such statement is desirable).

- ❑ **Data published** (Top)

OK, so now you've done it… Congratulations! :)

But the job isn't over yet, keep going on the list…

- ❑ **Documentation published** (Top)

Ensure your documentation is easily discoverable near the source of the data.

- ❑ **Feedback option available** (Top)

Also discoverable near the source of the data there should be an option for data users to provide feedback, either on the data itself or the ease/difficulty of access.

Ensure there is a process behind the feedback option so that you can be responsive and take action as required. Responsiveness to feedback can build trust and relationships that can be mutually beneficial, especially for future releases of data.

- ❑ **Invitation for reuse stories** (Top)

Don't be shy, ask users of the data to tell you how they are using the data. Explain that the more stories of use and impact from the data will help the business case for investing in more data releases.

You could even commit to publishing their story as a case study. This is often a win-win because it not only helps your case for investing in more open data, but is often good publicity for the data user and their business or advocacy/cause.

- ❑ **Listed on a central catalogue/portal** (Top)

Once you have published data, listing it on a central government data catalogue or portal can make it much more easily discovered and used.

For a single dataset this is often only a 10 minute job, providing a standard set of metadata about the dataset. If you are regularly releasing open data there may be ways to to bulk update or harvest your data listing.

Your central portal should have information about your options and how you go about updating the catalogue, for example at Data.govt.nz.

- ❑ **Availability promoted** (Top)

Don't just list the data and sneak away, tell the world about it. The more people know its there, the more likely people who have a need for it will become aware and use it. So tell any community you have built around your data, share on social media, newsletters, Meetups and conferences if appropriate.

And as people use it, gather the stories on how they have used it and share on all your channels.