Marine Geospatial Information stocktake guidance

NZMGI Working Group

Published in partnership between MPI, NIWA and Toitū Te Whenua LINZ

August 2021

Content

1 Context and background

In New Zealand, Marine Geospatial Information (MGI) is collected by various public and private organisations. Knowing what data exists is the first step enabling data accessibility and reusability and increasing the value of marine geospatial investments. Reduced duplication of collection and gap analysis for future investments can be achieved with discoverable data and notably through the creation of dataset inventories.

The NZ Marine Geospatial Information Working Group (NZ MGI-WG) is working to increase the value of NZ MGI by making it Findable, Accessible, Interoperable and Reusable (FAIR). In 2019 the NZ MGI-WG agreed a national MGI Stocktake/ Inventory as a first step towards achieving this goal, improving the findability of NZ MGI.

Taking stock of MGI comes down to inventorying metadata associated with the MGI (see Appendix 1). Upon completion, the MGI inventory is published and refers to all MGI data assets held by the organisation. In an initial phase of the National MGI Stocktake (the Stocktake), six organisations completed internal MGI stocktakes and publish their results as data inventories on www.data.govt.nz.

This initial phase of the Stocktake identified benefits, challenges and complexities in the process and highlighted not all MGI stocktakes are comparable. The learnings from the initial phase were used to develop supporting material. These guidance documents will support organisations to effectively plan and undertake an MGI stocktake in the future, by highlighting key areas for consideration and presenting possible options.

2 Benefits from undertaking a stocktake

A data inventory is designed to accelerate the building of a well-managed data infrastructure and provide greater transparency, visibility, and consistency of data held by an organisation. Organisations will receive direct benefits from undertaking a data stocktake, and there are also significant benefits to external stakeholders and the wider New Zealand Inc.

2.1 Internal benefits to the organisation

Internally, the organisation will increase data **findability** by:

- Understanding what data assets the organisation manages,
- Helping the organisations better manage, find, and reuse its own data,
- Minimising duplication across sector by sharing information with a consistent set of core attributes,
- Strengthening effectiveness of performing legislative responsibilities,

- Supporting data managers and administrators improve data management systems and processing, as well as scoping for data publishing needs given customer need and management overhead.

Improved **accessibility** will be achieved through:

- Prioritising the release of data based on knowing what is important to data users,
- Improving transparency, reputation and increase public trust and confidence in the organisation,
- Ensuring requests for information are directed to the correct organisation, saving time and money,
- Reducing effort of responding to Official Information Act requests (OIAs), as data is easily findable and accessible,
- Increased organisation's staff's satisfaction through mature data maintenance processes,
- Meeting customers' needs and expectations.

The organisation will improve data **interoperability** by:

- Improving its quality while reducing duplication and aging of redundant datasets,
- Promoting identification of Data Stewards and Data Custodians,
- Increasing efficiency and reduced costs of workflows and analyses,
- Improving data exchange,
- Reducing of data related risk to organisation.

Data **reusability** will be increased by:

- Enabling analysis of datasets by a larger pool of users,
- Increasing the value of the data (ROI),
- Assisting agencies maturity in best practice data maintenance (e.g. cyclic review of data currency, generation of metadata),
- Enhancing effectiveness and accuracy of decision-making,
- Providing better quality of products and services,
- Enabling accelerated opportunity for innovation,
- Heightened return of investment in data capture and maintenance,
- Improving income potential.

2.2 Wider stakeholder and NZ Inc. benefits

- Wider understanding of data availability across the sector (enable the public to search, find and understand data held by sector organisations, e.g. on data.govt.nz),
- Awareness of which organisation is responsible for the management of datasets,
- Improved data supply process (access published data directly),
- Reduce duplication in data collection,
- Encourage collaboration and partnerships,
- Increased opportunity for add value and innovation,
- Data reuse to support informed decision making,
- Opportunities to reach to a bigger group of external data users.

3 Preparing for the MGI Stocktake

Several organisations have undertaken their marine data stocktake. Following is a wide variety of learnings which can assist other organisations to mitigate challenges when preparing to undertake their own stocktake. The below section lists suggestions and advice to support a better stocktake journey.

3.1 Assess the current state

1. What marine data is the organisation responsible for? Can the data be mapped against MGI themes in Appendix 1?
2. Has the organisation undertaken a data stocktake in the past?
3. Who are the relevant data owners/managers within the organisation? You can refer to Appendix 2.
4. Where does the organisation data reside? Identify where the information is held by identifying the various systems your organisation is using. Various datasets, copies or versions of datasets can be held on different systems.
5. Are there different copies of the same datasets? These might be simple copies, different versions of a same dataset or distinct datasets (after the original collected data was processed).
6. What metadata does the organisation collect for marine geospatial datasets? Refer to Appendix 3.
7. Does the organisation have standardised naming conventions, schema formats and data vocabularies?
8. Does the organisation have a data catalogue or data management system?
9. What marine data is already publicly findable or accessible? And what channels/ platforms is used for publishing data?

10. Identify whether you want to take stock of all your data, or only those of high value for your customers, or only the processed datasets, etc. Define what high/low value means to your organisation / business team, what processed vs unprocessed dataset is.
11. What information can be reused from the above and what information needs to be captured?

➢ Points of interest

The Open Data Institute's **open data maturity assessment** (https://pathway.theodi.org/) can be used to establish a baseline for the organisation.

3.2 Plan the work

Data stocktakes can require more effort and time than initially expected. It is important to carefully plan and describe what the project will deliver and how this will be achieved so stakeholders have a common understanding and progress can be monitored and measured. The stocktake project plan should include:

- project deliverables
- scope:
  - priority datasets to be included
  - metadata details to be captured (see 4.1. Level of details)
- timeframes
- project approach (see Appendix 4),
- resources (also refer to Appendix 5),
- engagement/communication planning includes:
  - identifying key stakeholders and defining roles and responsibilities (refer to Appendix 2),
  - identifying approaches for communicating with the project team and across stakeholders,
  - associating stakeholders with focus areas (e.g. data management, publication/release),
- risks and issues management strategies,
- project management approach. Time-boxing the project will allow to maintain momentum. Implementing an agile approach (versus a waterfall approach) will provides benefits such as:
  - short sprint cycles (e.g. 1-week or 2-week cycles),
  - retrospectives at end of each cycle to demonstrate what was achieved and learnt with the product owner and/or project sponsor,
  - focus on the project for 6-8 consecutive weeks.

➢ Points of interest

Scope: Consider starting small, and prioritising data in one area, to create benefits for a specific business unit/area before scaling up.

3.3 Build internal support

Sponsor from the senior management team is vital for the prioritisation of the stocktake project and to ensure appropriate staffing resources are available. A well written proposal will assist the stocktake project getting underway more quickly:

- Clearly understanding and promoting stocktake benefits is crucial to achieve buy-in.
- A concise memorandum to stocktake sponsor (senior leader in the organisation) guarantees commitment from data stewards, ensuring required resources will be available (see Appendix 6).

4 Undertaking the stocktake

4.1 Level of details

The initial phase of the Stocktake uncovered different perspectives on the level of details (or granularity) captured in the different inventories. These discrepancies come down to differences in:

- understanding and agreeing on what constitutes a dataset (refer to Appendix 1),
- deciding on a relevant level of detail, i.e. at what level items need to be added to the inventory.

The use of lists of categories is recommended to ensure data quality and consistency for the inventory, and to contribute to a better user experience. It is left to the organisations to choose what level of details they want to capture their information. However, for the purpose of publishing the National MGI Inventory, we recommend that datasets are grouped under MGI themes and types (Appendix 1). Tags (using the proposed sub-types) can be added to the records to further describe what the datasets include.

4.2 Source of truth

Creating and using a single shared file will act as a single source of truth. Sharing multiple unique inventory files with stakeholders for their contribution and then bringing these contributions together afterwards tends to give rise to uncertainty around which version is the master version.

➢ Points of interest

Template: To support organisations creating their MGI inventory and contributing to the National MGI Inventory, we recommend reusing the NZMGI inventory template.

5 Publishing the stocktake

The value in undertaking a stocktake only comes from the created inventory being used, internally and/or externally. It is therefore recommended to develop communication to promulgate knowledge of the inventory's existence.

It is important to have a clear method of presenting the inventory to internal and external users. Although not all data discovered during the stocktake might be released to the public initially, its discovery can facilitate making datasets visible and encourage data access, eventually leading to the data being open. It can also encourage other organisations' specific Open Data initiatives (e.g. Environmental Monitoring graphs and data).

➢ Points of interest

The framework, processes and toolkit developed by NZ Transport Agency can assist organisations with releasing their MGI (https://www.data.govt.nz/blog/a/).

5.1 Licence and data release

Conflicting metadata can exist where a dataset is labelled as "open" while in fact it can only be released to internal staff. It is therefore recommended to ensure clarification with the business and subject matter expert (see Appendix 7).

Recommendations include:

- Ensure expected metadata attributes are completed to support publication,
- Find agreement or advice if metadata attributes cannot yet be "filled in".

5.2 Publishing options

If the MGI inventory is published on a website or portal, the link can be listed on the **NZ MGI website** to increase discoverability (seek support by emailing hydro@linz.govt.nz).

Public sector agencies can also adopt the free **data.govt.nz** catalogue service. To be published on data.govt.nz, the inventory needs to comply with core attributes set by the Department of Internal Affairs (DIA, see Appendix 8). The catalogues can be self-serviced and updated automatically via tools developed by DIA (see Appendix 8).

Organisations can also setup **data scraping** of their data portal to automate discovery of data through data.govt.nz (seek support from the DIA).

6 Maintaining the MGI stocktake

The maintenance of the inventory is critical to successful data management. It will avoid the inventory getting outdated and prevent the organisation taking stock of all the data again in the future. The various aspects of maintaining an inventory need to be considered.

6.1 Curating the existing inventory

At each scheduled maintenance review, the following high-level approach should be used:

- ensure new marine geospatial information has been identified and documented. This may include:
  - system queries to identify new files (i.e. created since the last review date);
  - consulting relevant staff to identify key recent projects / data collection efforts.
- remove any obsolete or superseded datasets from the inventory,
- review and update missing metadata, and ensure metadata standards have been adhered to (see Appendix 3),
- communicate and address any issues with information management or metadata standards that may have been identified during the maintenance review,
- ensure all staff have access to the updated content,
- update the public version on data.govt.nz and/or similar.

Ensure the maintenance process is defined (consider roles, technology, and infrastructure required), agreed, and supported by stocktake sponsor to keep the inventory up to date.

6.2 Refining the existing inventory

Another aspect of maintaining the inventory includes refining the catalogue. For this, we recommend that organisations publish their inventory. It is very likely that feedback from end-users will target data that would be most valuable to them. End-users may also require some records to be further detailed and split into different entries to facilitate discoverability of some information (see 4.1 - Level of details).

6.3 Expanding the inventory

Adding records to the created inventory as data is captured is one aspect of maintaining an inventory. The way metadata was extracted and what resources were used in the initial phase will likely dictate the process adopted to catalogue future data. Including recommendations and feedback from the team that participated in the initial stocktake will benefit the inventory maintenance plan.

Best practice metadata should be a focus of data management so that future records can easily be catalogued, particularly when relying on automated processes. Data standardisation will also provide consistent output. For instance, adopting naming conventions and vocabularies when collecting and cataloguing data will highly facilitate future data cataloguing.

When using bespoke scripts, changes to the scripts might need to be considered, if new attributes are recorded by new equipment/software or new types of data are collected in future surveys.

6.4 Version control

It can be useful to create version controls for updated inventories, especially if the inventory is openly published. An example of a name could be “Data inventory v1 20190910”. Should anything go wrong, specific versions of the inventory can be recalled, and the error or issue tracked down.

6.5 Update frequency

The existing inventory should be continually updated by any staff member regularly collecting or managing marine geospatial data, for example when data is:

- collected or obtained,
- significantly altered, updated or replaced,
- superseded, deleted or archived.

Establishing a plan to manage updates will allow for planning of resources and management of customer expectations. Updates can be set at regular intervals (monthly, quarterly) or match data collection (e.g. new updates every time a survey is undertaken or when equipment reaches storage capacity).

7 Appendices

Appendix 1: Definitions

Creating common terminology and definitions will help clarity for all stakeholders involved.

Stocktake, inventory and metadata catalogue

A **stocktake** is the process of inventorying data assets held by an organisation. Taking stock of digital information comes down to recording its metadata attributes into a database (i.e. inventory). An MGI inventory is therefore a collection (or catalogue) of metadata records (or entries or occurrences) related to MGI datasets.

Upon completion, the inventory should contain a comprehensive list of all data assets held by the organisation. This includes a set of core metadata attributes, consistent across government and non-government organisations, and, where applicable, a set of organisation-specific additional metadata attributes. More information on metadata is provided in Appendix 3.

MGI datasets

A **dataset** is often defined by data collection and data management practices (e.g. where data is collected over many years and results in many datasets, one record may be recorded for each dataset or datasets may be grouped as one record across the entire data range). A dataset is a structured collection of related data that forms a cohesive entity. It is part of the same management process whether it is the collection, QA, storage, curation, or publishing of the dataset. A collection of a related set of one or more data items is managed in the same business process and relates to a user theme. Data items are structured facts, statistics, observations or measurements collected together represented as text, numbers, geometries, or images. All data (or records) are collected/generated for a particular purpose, project, or using similar technologies. Datasets consist of one or more data items such as database tables, layers, or files, including the metadata. It is left to the discretion of the respective organisation to define the scope for “dataset”.

A **geospatial** dataset is a dataset where one or many of the items contain an explicit or implicit spatial attribute. This includes raster, grid, or vector layers, or data defined by an extent or point location.

Marine data relate to data with a marine component, existing in or produced by the sea, pertaining to navigation or shipping. This encompasses data collected from the coastal area to the deep ocean. Estuaries (such as bays, harbours, lagoons, inlets, or sounds) fit within this category as they form a transition zone between riverine and marine environments and are subject to marine influences such as tides, waves, and the influx of saline water.

Considerations

- The dataset purpose is not a defining characteristic for a dataset, and the dataset may be used for multiple purposes.
- The data items do not need to be the same file format, data type, or schema specification (e.g. file extensions/formats for bathymetric data can come in a number of different formats such as .all, .s7k, .db, .wcd, .qpd,.las/.laz, ASCII xyz, .csar, .tiff depending on how processed the data is).
- A database is an organised collection of data stored as one or more datasets stored and accessed in a computer system. So a database is not necessarily a dataset. A schema within a database is likely a container for a dataset.
- Different resolution of data should not dictate the dataset.
- Reference data that is used or referenced is not part of the dataset.
- Derived data might be part of the same dataset, which contains the raw data. However, this depends on how the data is identified by users.
- These characteristics are not absolute. In some cases, a dataset does not meet all the characteristics.

MGI themes

Categories and ontology recommended by the NZMGI-WG can be found here.

Appendix 2: Roles and responsibilities

Ambiguities within organisations can exist regarding role definitions for data management. Organisations that hold data from third-party suppliers need to identify the ownership of the datasets and clarify data that should be included in the stocktake. The initial phase of the Stocktake showed that decision was often made to only include datasets which the organisation was responsible for, and datasets of other agencies were excluded from the stocktake.

We found that various terminology was used interchangeably (e.g. owner, steward, maintainer, etc). It is therefore recommended to discuss individual roles and ensure agreed understanding has been reached before sharing definitions and assigning roles and responsibilities with all involved in the stocktake.

The following definitions were taken from the LINZ Data and Information Management Policy 21 February 2020 and the Data Toolkit published by data.govt.nz. Definitions were extended by the NZMGI-WG and are proposed for use when undertaking an MGI stocktake.

Custodian, the ‘Owner’

Role

The Department of Internal Affairs refers to data custodian as the person within the organisation responsible for updating, maintaining, and preserving the data. Under the Well Managed principle, agencies, as the stewards of government-held data and information, must provide and require good practices which manage the data and information over their life cycle, including catering for technological obsolescence and long-term preservation and access.

Responsibilities

- Ensuring that organisation’s data is accessible and well maintained
- Act as the contact person for queries or requests about the data

Domain Steward, the ‘Guider’

Role

A Domain Steward works at a strategic level, takes a holistic view, and operates within the context of the organisation.

They set the strategic direction for the data domain, advocate for the needs of data users at the macro level and use their influence to promote good practice through their awareness and understanding of the statutory and regulatory environment.

Responsibilities

- Setting the strategic direction for the data domain
- Designing, setting and monitoring high-level frameworks and standards
- Engaging externally to influence and improve the quality of data and information sets strategically important to, but not held by, the organisation
- Promoting strategic investment in, and use of, the data by articulating benefits and opportunities
- Monitoring and adjusting strategy and standards
- Facilitating the appointment of the Data Manager
- Managing the following risks:
    - Change in Government policy (risk/opportunity)
    - Economic factors
    - Data Manager and Data Maintainer roles not assigned

Data Manager, the 'Controller'

Role

A Data Manager defines the appropriate operational data/information policies and standards and provides/secures the funding and resources for a data maintainer to effectively manage and deliver datasets to the users.

Typically, a Data Manager will be at Team Manager level or above and will have financial and people management delegations. The specific level depends on the size and complexity of the dataset. They receive their mandate from a Deputy Chief Executive.

Responsibilities

- Bidding for, obtaining, assigning and overseeing required resources (including funding and people)
- Aligning to the strategies for the data domain defined by the Data Steward
- Defining the operational framework for the information/data lifecycle
- Owning data risks and setting controls

Data Maintainer, the 'Doer'

Role

This role administers a data or information set throughout its lifecycle by implementing the operational data/information policies and standards defined by the Data Manager. While accountability for implementation rests with the Data Manager, they may delegate responsibilities to others.

Responsibilities

Operational management of data/information through its lifecycle, following the framework defined by the Data Manager. This includes activities specified in the data and information lifecycle.

Data Users

Role

Data users from the user community within and external to the organisation, play a key role in ensuring that the policies, requirements, delivery mechanisms and technical architecture designs meet the needs of the user community. Without the data users there would be no need for information to be governed. Data users must understand their information needs sufficiently to assist in the data governance function and comply with the relevant data and information management use responsibilities.

Responsibilities

- Comply with all relevant organisational policies, standards and operational processes regarding the creation and use of organisation's data and information throughout its lifecycle.
- Provide feedback on data and information requirements to Data Managers to inform the ongoing development and management of data and information sets.
- Engage with Data Managers to facilitate improvements to existing data sets and the creation of new ones, through such processes as co-creation or collaboration.

Stocktake champion

Role

Nominating a stocktake champion is highly recommended to help the success of the stocktake. The organisation's data champion needs to have a genuine interest in the creation and use of the inventory. The role of the stocktake champion is to lead the stocktake by:

- Acting as key point of contact across the business group,
- Ensuring that the inventory is consistent across the different types of data and across the different business teams.

Responsibilities

- Creating buy-in from internal stakeholders
- Being an escalation point if subject matter experts' responsiveness is low

- Providing peer review and subject matter expertise
- Navigating the internal networks (finding and liaising the appropriate staff)
- Being the conduit between business and contractor undertaking stocktake (if external help is being contracted in)

Appendix 3: Metadata guidance

The metadata guideline can be found here.

Appendix 4: Approaches for capturing the metadata

Capturing or extracting metadata can be done manually or automatically. The choice of manual or automated approaches depends on various criteria (i.e volume and complexity of the data, data management practice, available resources) that will need to be examined by the organisations. A combined approach should not be ruled out, as automated and manual options can complement each other. This combined approach should be based on costs and benefits of desired outcomes.

Manual approach

The manual approach to capturing metadata involves staff extracting information (*refer to Metadata standard doc*) from folders and files and recording the information in a catalogue.

Certain situations are better suited to a manual approach. These include:

- where there is a lack of consistency or completeness in recorded information and data management e.g. inconsistent naming conventions
- small number of records to be captured.

The manual approach allows for checking files/folders individually if required. Staff may need to look more deeply at datasets if existing names/metadata are not specific enough to extracted required metadata details. This allows metadata to be captured with high confidence, despite the possibility of human error.

Subject Matter Experts (SMEs) may be required to provide input depending on dataset complexities, technicalities, data management practices. Involving a SME may reduce the risk of mis-classifying data.

Additional benefits of the manual approach include:

- collaboration across teams, when data has been collected for various projects and managed by different teams.
- identifying opportunities to improve data management by picking up.

Automated approach

During the initial phase of the Stocktake, initial full scans of data disks undertaken by organisations recorded tens of thousands of distinct items of spatial data held on personal and departmental file shares. A cursory check had not revealed what percentage of this unmanaged data was of value for possible reuse. Unassisted automated scanning does not provide insight into usefulness of discovered datasets or files but can indicate the scale of potential records to be assessed.

Automated approaches have the advantage to be time efficient especially when the amount of data is substantial, and the processes can be repeated across different systems. They provide consistent outputs that can directly feed into the inventory and generally require limited human input if well designed.

Information extracted by automated tools is however limited to the metadata associated with the file (e.g. name, format, size, date of creation, etc.) or stored in a database. Understanding how the tool operates (including the set criteria) and what information is captured by the tool are essential prior to designing and/or using the tools. Failing this will result in not-fit-for-purpose outcomes.

The efficiency of automated tools and their outcomes depends on good data management practices and the quality of selection criteria. If the formats of the files, the naming conventions or any other set criteria are not consistent or change over time, automated tools (especially scripts) will provide incorrect and/or incomplete outcomes. Manual human quality checks of random files can help validate and refine the processes.

Several options exist when using automated tools to extract metadata information.

a) **Data scraping** refers to extracting information from various sources, including a local machine, a database or the web. "Off-the-shelf" tools exist and typically include a crawl agent and parser. Crawling is the process of navigating through the system (referred as web crawler when searching through data published on the internet) and fetching its content. Scraping extracts specific data from the content which has been fetched. Parsing takes the fetched document, parses it and extracts required information from it.

b) **Bespoke custom scripts** can be developed (using python for example) to identify datasets using a set of criteria (i.e. file names or key words, format, size, and spatial metadata). File naming conventions enable the script to identify and classify data.

c) **Data catalogue software** includes metadata management tools that inventorise, organise, and maintain inventories of available data assets within systems. The inventory combines metadata from various systems into a catalogue where relevant data is discoverable. Besides relying on traditional metadata management, additional features can include glossaries that can be mapped to specific data assets, data lineage documentation, or even more advanced features to support business analyst teams.

Dedicated resources are required when using automated approaches. This will be in the form of purchasing "of-the-shelf" tools or resourcing expert staff (i.e. developers and SMEs) to develop and validate bespoke tools.

Summary of approaches

Focus	Options	Strengths	Weaknesses	Risks	Dependencies	Opportunities
Capturing/ Extracting metadata	Manual approach	• Detailed investigation • Confidence in deliverables	• Possibility of human error • Resource consuming	• Misclassification of records • Inconsistent records if multiple inputs	• SMEs knowledge (depending on data complexity)	• Identification of data management issues • Internal collaboration
	Automated approach	• Wide focus/target • Thorough investigation through the systems • Time efficient • Repeatable • Consistent output • Relative limited human input	• Expenses related to the purchase of the tools or to resourcing staff developing the tools • Limited output in terms of metadata information	• Limited outcomes if poorly designed tools or poorly managed data	• Understanding of tool process and requirements • Dependent on good data management practices • Dependent on set criteria	• Can be used in isolation on specific datasets or systems.

Appendix 5: Resourcing the work

The effort required to undertake an MGI stocktake is dependent on the desired output, the amount of data, and the maturity of data management within the organisation. A stocktake can be a significant undertaking and committed resources are required to see it through.

Internal resources

Many MGI datasets are very specialised and have unique characteristics and complexities. There may be broad range of different MGI datasets with an organisation, with SMEs across various parts of the business. Using internal resources who are familiar with the data, their formats, and where the data is stored, may result in higher quality, more comprehensive output.

Internal staff have existing and direct access to the systems and databases, reducing security risks and policy breaches. Finally, resourcing own staff will leverage internal skills and increase cross organisational collaboration.

External consultants

An organisation might choose to engage a consultant to undertake the stocktake on their behalf. This option was offered during the initial phase of the Stocktake by Statistics NZ within the Open Data framework.

This option has the benefit of relying on dedicated professional resources who are experts in data asset management. While this option requires budget, there is a high degree of certainty around deliverables, timeframes, and costs.

Consultants have access to specialised tool and techniques that will be adjusted according to the amount and complexity of the data.

In this model, most of the work falls on the consultant, freeing staff time but limiting input around deliverables or opportunities to refine processes as data is uncovered. SME will still be expected to assist, especially when data is complex, storage systems varied and formats uncommon.

If this approach is chosen, the organisation will need to consider opening their system to the consultant. This method will likely need to involve the procurement team and in some cases the HR and IT teams as well.

Additional learnings include:

- Ensure clear communication to engage data stewards/other stakeholders with consultant.
- Consistent internal data stocktake champion and consultant for duration of stocktake. Synergies and mutual understanding if not documented has potential to delay or derail the project,

- Project management function required (guiding stand-ups between champion and consultant, agreed milestones, reporting expectations, etc),
- Regularity of stand ups to support momentum,
- Develop clear understanding of scope, definitions and processes to ensure delivery of expected outcomes,
- Define and agree consultant's responsibilities clearly, e.g. will data steward or custodian confirm the open/not open status or is consultant assigned to check metadata and other reference material,
- Quality expectations are clearly communicated, e.g. include only organisation's 'own' data, ensure correct status (open versus not open), check any URL-links.

Summary of resource options

Focus	Options	Strengths	Weaknesses	Risks	Dependencies	Opportunities
Resourcing	Internal staff	• Familiar with data, formats, systems, and storages • Knowledge of hidden resources (unrecorded metadata, unvalued data stored "elsewhere") • Understanding of the value of the data • Existing access to data and systems	• Competing organisational priorities	• Comprehensive knowledge can affect overall outcomes	• Availability of staff/SME	• Collaboration across teams • Leverage internal knowledge • Educate on data management practices
	External consultant	• Limited input from SMEs • Access to professional expertise and specific tools • High confidence around timeframes, costs, and outputs	• Reimbursement costs • Limited control over the process/output • Project only covers what is available/visible to the consultant	• Understanding of data generally limited to non-complex data • Security risks associated with opening systems to consultant	• Consultant availability and understanding of the data • Procurement process involving proposal and contract • Physical and/or digital access to the organisation and/or system • SME input for specific and/or complex data	• Professional approach to data management • Professional assessment of information asset • Professional education f internal staff

Appendix 6: Sponsorship templates for undertaking the stocktake internally

Internal memorandum to Director <*organisation*>

From:	<*Champion name, role*>
CC:	<*Champion support name, role*>
Contact:	(02x) xxx xxxx
To:	<*Sponsor name, role, organisation*>
Date:	dd month year

Subject: <Organisation> Data Inventory

Purpose

1. This memo seeks your support for the <*organisation*> Data Inventory project.

Recommendation

a. **Note** that this <*organisation*> Data Inventory project is a continuation of an earlier successful pilot (<*date*>) [*insert if applicable*]

b. **Note** that this project supports a national programme of work started by the New Zealand Marine Geospatial Information Working Group (NZMGI-WG).

c. **Note** there are no capex cost implications for <*organisation*> related to this project [*This recommendation assumes that all work will be undertaken by internal resources.*]

d. **Note** the project's benefits for both <*organisation*> and the public are detailed in this memorandum

e. **Agree** to:

   i. Communicate this project's intent with your leadership team and

   ii. encourage your leadership team to make their key staff available when required

Key Points

2. This <*organisation*> Data Inventory project is a continuation of the earlier pilot from <month> <year> which, due to <budget constraints, limited resources…. delete what not applicable, adjust according to the reason>, had not covered <*organisation*> entire datasets. [*insert if applicable*]

3. The <*organisation*> Data Inventory project is planned to commence day month year.

4. The overall NZMGI work programme initiated a public facing Marine Geospatial Data Inventory hosted on Data.govt.nz. This <*organisation*> Data Inventory project will support this initiative by providing metadata about listed <*organisation*> data assets to allow for considerations on appropriate re-use. The inventory will as well inform whether the data is available for re-use, unavailable, or awaiting assessment under the <organisation's> open licence policy. If data is open and accessible, a direct link to the data will be provided in the inventory.

5. This project is aligned with the <*organisation*> Spatial Data Audit initiative currently underway and managed by <*organisation*>*'s* GIS Team. *[if applicable*]

6. The main purposes for undertaking this project are that:
   a. It will help the public discover what <*organisation*> data is being held and available
   b. It will help <*organisation*> to better understand what data assets we hold and are responsible for managing and maintaining
   c. A cross government and industry data inventory will help inform data sharing and reduce duplication of data

7. It should be stressed that there is no requirement for <*organisation*> to list every data asset it holds on the inventory. Part of the information captured by the project team will help inform whether publishing information about a data asset will have any negative repercussions and decisions about what is made visible will be made accordingly.

<Organisation>'s involvement

8. The project team will build upon the metadata captured in the existing <*organisation*> enterprise data catalogue and spatial data repository maintained by <*organisation's GIS team*> and will include any additional information that is recommended by the New Zealand Data Inventory programme guidelines.

9. This work will support the refresh of the existing <*organisation*> enterprise data catalogue as the internal and external inventories will be based on the same central source with different attributes exposed depending on the audience.

10. Upon completion of the stocktake, the project team will report back to you with the completed list, including a clear indication of what datasets could be listed on data.govt.nz.

11. The project team will deliver a process documentation to support ongoing maintaining the inventory. The documentation will flag challenges to improve <*organisation*>'s data management.

12. Statistics New Zealand on-line resources[1] are available to provide <*organisation*> with advice and guidelines for this inventory, with the intention that this will equip <*organisation*> with the tools and knowledge to produce a completed inventory if required.

13. Guidelines from the NZMGI-WG, including benefits, processes and metadata guidelines will support this stocktake.

Resource Implications

14. It is expected that resource requirements from <*organisation*> staff will be moderate.

15. The resource will interview key staff about the data they collect, maintain, and use and will collate the relevant metadata throughout discussions. It is anticipated that interviews will last approximately between 30-60mins person and, after the contract resource has collated the relevant information, staff will be asked to validate the metadata that has been captured before any inventory register is published. *[include if a statement on high level process details is required*]
16. <*Organisation*>'s staff <*name, role*> and <*name, role*> are co-leading this project ("Project Team").

17. This work will inform the refresh of <*organisation*>'s internal enterprise data catalogue with the aim that <*organisation*>'s staff contributions will be small: once for initial input and then for final review.

[1] Fact sheets: https://www.stats.govt.nz/assets/Uploads/Data-leadership-fact-sheets/Fact-sheet-open-data-Mar-2018.pdf
Open Data definitions: https://www.digital.govt.nz/standards-and-guidance/data/open-data/
Open government information and data programme reference: https://www.data.govt.nz/open-data/open-data-nz/

Benefits

18. Completion of a publicly accessible inventory will result in a number of benefits for both the public and <*organisation*>.

19. Benefits for the public

- Increased external visibility of <*organisation*>'s data assets
    - Allowing the public to link directly to already open and accessible datasets makes data more easily discoverable and accessible
- The public can easily see if data has been released, assessed or is not able to be released
- The public can easily see which agency holds a particular dataset and can direct their request to the correct location
- The inventory will provide context on the data that <*organisation*> holds, and the potential for reuse

20. Benefits for <organisation>

- Increased internal visibility of <*organisation*>'s data assets directly feeding into <*organisation*>'s Data catalogue
- Improve transparency and increase public trust and confidence in <*organisation*>
- Increased external visibility of <*organisation*>'s data assets [*add the following for governmental organisations:*]
    - Allowing the public to link directly to already open and accessible datasets will reduce the number of OIAs requesting data that is already open and available.
    - Supporting <organisation> as part of an "Open and Transparent Government"
- Key tool for prioritisation of data release
    - Requests for data 'not yet assessed' will help us understand what data the public are interested in,
    - Candidates for proactive release will be discoverable based on repeat requests.
- Reduced requests for data that is not held by <*organisation*>,
- Meet customers' needs and expectations,
- Meet our legislative requirements, *[if applicable*]
- Ensure requests for information are directed to the correct agency, saving time and money,
- Providing <*organisation*> with a broader picture of the data assets that it holds and will increase the emphasis on information and data management practices as data becomes more visible outside of the agency,
- Opportunity to utilise the resources that are provided to begin to refresh our own enterprise data catalogue at no cost,
- By increasing the visibility and potential release of data from the <*primary industries/education/research and science community* [*delete non applicable category*]>,

the <*organisation*> has the potential to benefit from the data being used in innovative and unexpected ways to enhance primary industries operations and activities,

- Better understand the value of our data,
- <*Organisation*> has multiple publishing channels for geospatial data and many potential data stores which are not published. Some of these channels include data that is not either owned by <*organisation*> and <*organisation*> publishes the data on behalf of other agencies. Having a clear understanding of <*organisation*>'s holdings will support data managers and administrators to scope for future data publishing needs, determine which open data channel and capability are appropriate given customer need and management overhead. *[if applicable*]

21. Benefits for <*organisation'*s sub team(s)> and wider organisation

- The inventory will result in a comprehensive list of data assets within the <*organisation's specialised subject matter*> system,
- Increased visibility for internal staff and external stakeholders on data assets and improved transparency,
- Improved operational efficiency,
- Making it easier for external parties to access data for their day-to-day operations,
- Increased emphasis on data management practices will enable improvement in data quality,
- Better information on data assets could assist with future system performance improvement projects, such as the governance framework review and service improvement initiatives,
- Reduced effort by <*organisation*> and <*organisation*>*'s* GIS Team staff to respond to data requests,
- Supporting data administrators and data managers and improving data management systems and processing via a better understanding of data holdings.

Appendix 7: Guidelines and references for data publication

The following criteria are to assist data providers decide what data can be published and under which terms. It also assists data users in what needs to be considered when using the data. The guideline is based on the:

- Declaration on Open and Transparent Government. https://ict.govt.nz/guidance-and-resources/open-government/declaration-open-and-transparent-government
- New Zealand Government Open Access and Licensing (NZGOAL) framework. https://www.data.govt.nz/toolkit/policies/nzgoal/
- NZTA framework and process for opening data https://www.data.govt.nz/blog/a/

What to publish?

- Who is the authoritative source of the data? Who is nominated custodian of data?
- How is collection, management, and dissemination of data funded? If government funding is used then under the Open Government Declaration, data should be made available with as little as possible restrictions for re-use.
- Are there any Intellectual Property rights attached to the data?
- Is any of the data of sensitive or personal nature?
- What risks are related to make the data publicly available? (E.g. bio hazards; risk of ecosystem degradation, etc.)

How to publish?

It is best practice to publish the data under Terms / Conditions / Licences. These can protect the data provider and should consider the above points on ‘what to publish’.

How to use?

Data users need to recognise and act upon the conditions (licence) under which the used data has been published. This includes, but is not limited to considerations like the following:

- If the terms / licence for a data source publication includes a ‘by attribution’ (BY) clause, any re-publication of the data or publication of a product that is based on the data shall include a reference to the data source and data provider. We suggest it is best practice to acknowledge the data source and data provider in any case of related publication.
- If the terms / licence for a data source publication includes any type of waiver of responsibility or data limitation the data user shall not, directly or indirectly, infer any responsibility for consequences of data re-use or publication to the data provider.
- If the terms / licence for a data source publication includes a ‘non-commercial’ (NC) clause, this data source shall not be used for any commercial application.

- If in doubt about anything with regards of re-publishing data or related products, contact the data provider.

Appendix 8: Publishing on data.govt.nz

Set of core attributes

The following are core attributes required for inventories to be published on data.govt.nz:

Data field	Required	Example value
title	Yes	New Zealand Public Sector Websites
description	Yes	List of websites owned and administered by the New Zealand Public Sector. The Department of Internal Affairs acknowledges this list has been compiled to the best of their knowledge, but it is not a complete list of all Public Sector websites. This list will be updated as the Department becomes aware of required updates.
identifier	Yes	https://webtoolkit.govt.nz/guidance/domain-names/new-zealand-public-sector-websites/
Is Open	Yes	Yes / No
Can be open		Yes / No
What is required to be open ?		e.g. Private impact assessment, aggregation anonymisation, etc.
Estimated date of release		5/02/2000
licence	Yes	https://creativecommons.org/licenses/by/4.0
keywords	Optional	websites, open government, url
issued	Optional	26/08/2011
modified	Optional	1/04/2015
publisher.name	Yes	Department of Internal Affairs

Data field	Required	Example value
publisher.mbox	Yes	info@dia.govt.nz
contactPoint.fn	Yes	Jane Doe
contactPoint.hasPhone	Yes	4123456789
contactPoint.hasEmail	Yes	contact@agency.govt.nz
landingPage	No	https://webtoolkit.govt.nz/guidance/domain-names/new-zealand-public-sector-websites/
updateFrequency	No	Annual
theme	No	Fiscal, tax and economics
temporal	No	2011-08-26/2015-04-01
spatial	No	{"type":"Polygon","coordinates":[[[165.6298828125,-47.5468715989],[179.3408203125,-47.5468715989],[179.3408203125,-33.9068955513],[165.6298828125,-33.9068955513],[165.6298828125,-47.5468715989]]]}
distribution.0.downloadURL	Yes	https://webtoolkit.govt.nz/files/PublicSectorWebsites01April2015.csv
distribution.0.title	Yes	Snapshot at 15 March 2017
distribution.0.description	No	
distribution.0.format	No	CSV / text
distribution.0.size	No	58Kb
distribution.1.downloadURL	Yes	https://raw.githubusercontent.com/GOVTNZ/public-sector-websites/master/public-sector-websites.csv
distribution.1.title	Yes	Latest dataset updates
distribution.1.description	No	
distribution.1.format	No	CSV
distribution.1.size	No	68Kb
distribution.2.downloadURL	Yes	https://www.govt.nz/api/v2/consultation/list
distribution.2.title	Yes	A list of consultations
distribution.2.format	No	API
distribution.2.size		

Publication process

Organisations publish and update their inventory by emailing DIA directly (info@data.govt.nz).

Organisations can also take ownership of this process by generating a data.json file that will be harvested into data.govt.nz. For this follow the steps:

Prepare for pre-release

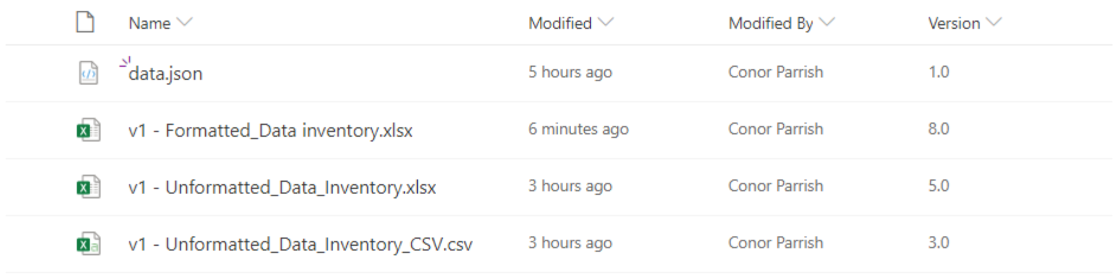
1. Find out where to store the json file in the Content Management System.

Check with your organisation's Content Management team where to store the json file in the Content Management System (CMS). The location of the stored file will produce a publicly accessible URL.

The URL structure can be https://YOURORGANISATION.govt.nz/data.json, or https://ird.govt.nz/media-library/tax-statistics. This URL link will need to be specified when using the json converter.

2. Create the folder structure for the inventory files in an easily accessible shared location.

This will need to incorporate a hierarchy of folders to ensure that the data inventory files are stored consistently and not mixed up, as the inventory is updated. The image below illustrates the files that will be kept within the v1 inventory folder.



Name	Modified	Modified By	Version
data.json	5 hours ago	Conor Parrish	1.0
v1 - Formatted_Data inventory.xlsx	6 minutes ago	Conor Parrish	8.0
v1 - Unformatted_Data_Inventory.xlsx	3 hours ago	Conor Parrish	5.0
v1 - Unformatted_Data_Inventory_CSV.csv	3 hours ago	Conor Parrish	3.0

3. Ensure no other existing file named 'data.json' is saved on your computer, otherwise your computer will save the file as 'data(2).json' by default, and this name will not be identified by the script if you upload this file to the CMS.

Therefore, to save yourself from renaming the file every time, we recommend that you simply delete the existing 'data.json' file from your local drive, once it has been uploaded into both Stax and the CMS.

Create the j.son file

1. The original formatted file of the inventory should be saved and maintained for consistency and information management purposes. Save the file as '[version number] – Formatted_Data_Inventory.xls'. For example, v1 – Formatted_Data_Inventory.xls . This file is the source of truth.

2. For publishing purposes this inventory file needs to be unformatted and saved in a xlsx format as '[version number] – Unformatted_Data_Inventory.xls. For example, v1 – Unformatted_Data_Inventory.xls. To unformat:

   a) Highlight the whole sheet and select 'Normal' (Home tab) to remove all colour fill, borders and other formatting that was included for readability.

   b) Highlight the column headers, go to the 'Data' tab and select the 'Filter' button to remove the dropdown filters.

3. Convert and save the file as a .csv UTF8 file called '[Version Number] Unformatted_Data_Inventory.csv. For example, v1 – Unformatted_Data_Inventory.csv

4. Convert the .csv file into .json by using either of the following tools (they both look and operate the same way):

   a) downloading the open source nodejs convertor tool (nodejs convertor) produced by Stats NZ, or

   b) running the .csv file through the online prototype convertor tool found at https://datagovtnz-csv-to-dcat-json.herokuapp.com/.

How to download and install the open source nodejs convertor tool?

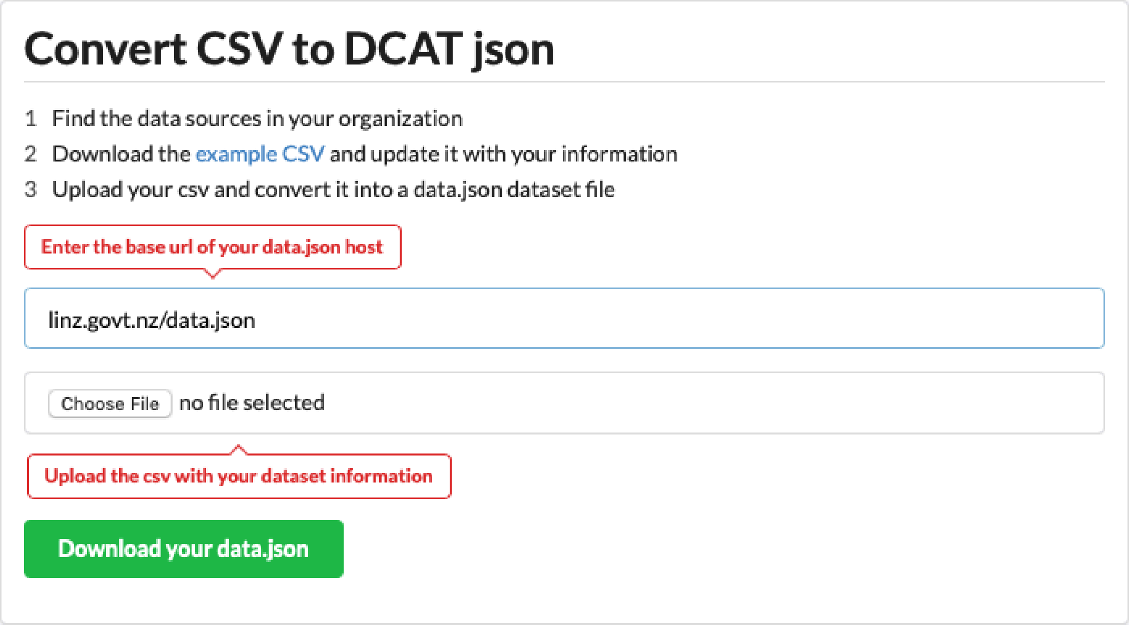
The source files for the nodejs convertor can be found on GitHub at: https://github.com/data-govt-nz/schema

Here are the steps, outlined in the 'README' master file on Github:

1. Install node.js and the npm package manager which you can get at https://nodejs.org/en/download/
2. Using the command line or other tool, run the node install command, this will install any other modules and related dependencies required to run the data.json conversion tool.
3. Navigate into the root directory of this code and run the following command to perform the conversion: node convert.js --url https://www.YOURAGENCY.govt.nz --file /PATH/TO/FILE/datasets.csv --output /PATH/TO/DIRECTORY
   - --url: your agency website address.
   - --file: the path to your CSV stocktake file.
   - --output: the path to where the resulting data.json will be saved, ensure you include the tailing / on the end of the path

How to convert the .cvs to .json using the data.govt.nz convertor tool?

1. Navigate to http://datagovtnz-csv-to-dcat-json.herokuapp.com/
2. Enter the base URL of your data.json host. This is the path to the location of the json file; ideally the file should reside at your organisation's publicly accessible website (e.g. https://ird.govt.nz/media-library/tax-statistics/data.json).
3. Upload the .csv file prepared earlier and select the 'download your data.json'



Convert CSV to DCAT json

1 Find the data sources in your organization
2 Download the example CSV and update it with your information
3 Upload your csv and convert it into a data.json dataset file

Enter the base url of your data.json host
linz.govt.nz/data.json
Choose File no file selected
Upload the csv with your dataset information
Download your data.json

4. Save the file in the shared location you have created during the pre-preparation steps.

To ensure the conversion was successful, you can validate the .json file here https://jsonlint.com/. If you would like to view this file, you can download a text editor application for this purpose such as atom.io or notepad++.

Save the data.json file in the Content Management System

Send the data.json file to the Content Management team to upload and publish the 'data.json' file to the CMS location, as defined above. The exact steps involved in publishing the data inventory to the CMS will be defined by the internal processes followed by the Content Management team. An overview of the inventory management process is summarised below.

Initial Publication:

1. Remove formatting and convert xls inventory file to csv.
2. Upload csv file to node.js converter.
3. Download data.json file.

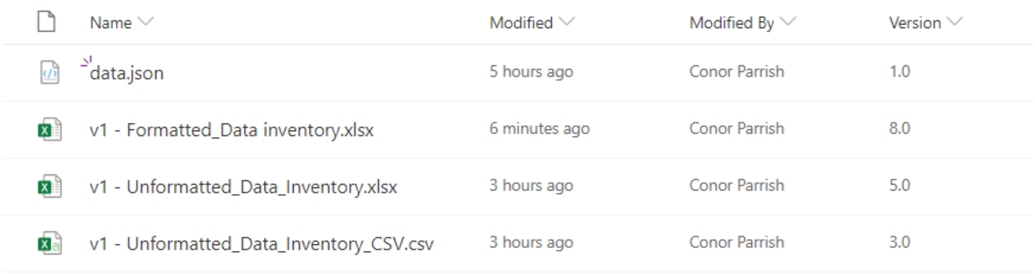
4. Upload data.json file to the shared location.
5. Upload data.json file to CMS.
6. Delete the data.json file from the computer.

Subsequent Publications:

For all subsequent updates, there is a slight variation from the initial publication process, necessary to maintain both a current ‘data.json’ file in the CMS, and a complete record of the evolution of the data inventory in Stax. These maintenance steps have been listed below.

1. Outdated copies of the inventory files are archived in the v1 folder.

    a. Rename the ‘data.json’ file in Stax as ’v1 - data.json’

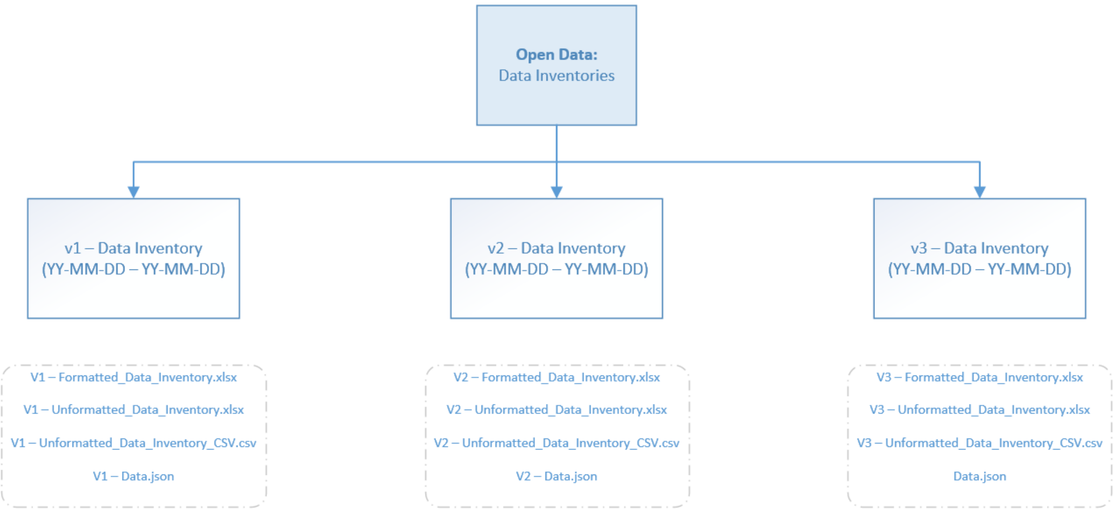


Name	Modified	Modified By	Version
data.json	5 hours ago	Conor Parrish	1.0
v1 - Formatted_Data inventory.xlsx	6 minutes ago	Conor Parrish	8.0
v1 - Unformatted_Data_Inventory.xlsx	3 hours ago	Conor Parrish	5.0
v1 - Unformatted_Data_Inventory_CSV.csv	3 hours ago	Conor Parrish	3.0

    b. Copy the ’v1 – Formatted_Data_Inventory.xlsx’ file into a new folder named: ’v2 – Data inventory YY-MM-DD – YY-MM-DD’.

2. Make the required updates to the file, then save the file as: ’v2 – Formatted_Data_Inventory.xlsx’
3. Remove formatting and convert the xls inventory file to csv (saving the files in the v2 folder).
4. Upload the csv file to the node.js converter,
5. Download data.json file.
6. At this point, there could be a data.json file existing on the computer, if there is an existing data.json then the download will be called data(2).json, and you must rename this file to ‘Data.json’
7. Upload the data.json file into v2 folder in Stax as ‘data.json’.
8. Upload the data.json file in CMS.
9. Delete the data.json file from the computer.

A potential file structure to manage updated versions of the data inventory could look like:



Open Data:
Data Inventories
v1 – Data Inventory (YY-MM-DD – YY-MM-DD)
v2 – Data Inventory (YY-MM-DD – YY-MM-DD)
v3 – Data Inventory (YY-MM-DD – YY-MM-DD)
V1 – Formatted_Data_Inventory.xlsx
V1 – Unformatted_Data_Inventory.xlsx
V1 – Unformatted_Data_Inventory_CSV.csv
V1 – Data.json
V2 – Formatted_Data_Inventory.xlsx
V2 – Unformatted_Data_Inventory.xlsx
V2 – Unformatted_Data_Inventory_CSV.csv
V2 – Data.json
V3 – Formatted_Data_Inventory.xlsx
V3 – Unformatted_Data_Inventory.xlsx
V3 – Unformatted_Data_Inventory_CSV.csv
Data.json

As part of the publishing process on data.govt.nz, the organisation will need to decide when the inventory is harvested (or updated) on data.govt.nz. This process involves extracting the contents of the data inventory from the json file to automatically re-populate data.govt.nz with the latest dataset updates. The frequency of when this process occurs can be set to daily, weekly, monthly (every 30 days), yearly or on an ad hoc basis.