3 APPROACH FOR RELEASE – STEP 3 OPEN DATA PROCESS

Data needs to be in a format that makes it easy to use, transform and reuse. It is important for the community to have confidence that the data they are accessing is current, reliable and well managed.

We will favour approaches that securely automate the release of regular or live open data services direct from our systems in order to generate sustainable value.

It is the Data Authorities responsibility to ensure a sustainable approach to the release of data is developed and approved.

It is recommended that the Data Manager of the dataset is engaged to advise and recommend the most suitable approach for the release of data. The Data Manager will have primary accountability for the day-to-day management of the information systems where the data is managed.

The approved approach will commit the agency to ongoing support and management of the released data in the format, frequency and delivery method determined. For this reason it is recommend that decisions in this section are included for Approval by the Open Data Authority and Advocate.

The approach for releasing your data includes:

- the format(s) the data will be released in
- the frequency of release
- the method of delivery
- consider data distribution services
- determine metadata and supporting information
- funding open data investment (automated delivery methods)
- cost recovery (cost recovered data is not in scope until further policy is developed).

3.1 FORMAT

For data to be open, it must be released in an open format and machine readable.

Open format: An open format is one, which is platform independent or non-proprietary, which means it does not matter what operating system or licensed software you have access to.

Machine readable: means that a computer without human aid can read data both in its format and in structure. Machine readable data is structured and easy to query using software code. Developers can consume the data from their programs/applications and re-use it.

Data may need to be transformed into another format to make it open. Releasing data in an open format may be as simple as saving or exporting system generated data in csv format instead of Microsoft Excel or PDF which is proprietary. Alternatively tools may be required to convert the data.

To understand what open formats are available refer to the Table 2 Examples of Common Open Data Formats.

Table 2: Examples of Common Open Data Formats

Format Name	Definition	Type of data to use this for
Comma Separated Values (CSV)	Comma Separated Values (CSV) is a great way of storing large amounts of data with just commas separating the data values. Often the csv file will contain a header with names describing what data is populating the file.	Tabular data e.g. Use instead of Excel
Tab-Separated Values (TSV)	TSV is a very common form of text file format for sharing tabular data and is highly machine readable.	Tabular data Use instead of Excel
JavaScript Object Notation (JSON)	JSON uses human-readable text to transmit data objects consisting of attribute–value pairs. It is used primarily to transmit data between a server and web application, as an alternative to XML. The file size will be more compact or smaller than XML	Complex structured data Multidimensional data Tabular
Extensible Markup Language (XML)	XML is a widely known markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. Users create and define their own tags.	Complex Structured data Multidimensional data Tabular data e.g. database extract metadata
Rich Site Summary (RSS)	RSS (originally RDF Site Summary), often dubbed Really Simple Syndication, uses a family of standard web feed formats to publish frequently updated information: blog entries, news headlines, audio, video. An RSS document (called "feed", "web feed" or "channel") includes full or summarised text, and metadata, like publishing date and author's name.	Use for announcements or events e.g. on websites
ATOM	The Atom Syndication Format is an XML language used for web feeds. The Atom format was developed as an alternative to RSS. Note RSS is the preferred standard	Use for announcements or events e.g. on websites
Open Document Format for Office Applications (ODF)	The Open Document Format for Office Applications (ODF), also known as OpenDocument, is an XML-based file format for spreadsheets, charts, presentations and word processing documents. It was developed with the aim of providing an open XML-based file format specification for office applications.	Non-system generated metadata or additional information you release with your dataset. (replaces Excel, Word, PDF)
HTML	Used for formatting information on the web	Non-system generated metadata or additional information you release (replaces PDF, Word)
Keyhole Markup Language (KML)	KML is an XML language focused on geographic visualization, including annotation of maps and images.	Spatial/location data
Geography Markup Language (GML)	GML is the XML grammar defined by the Open Geospatial Consortium (OGC) to express geographical features. GML serves as a modelling language for geographic systems as well as an open interchange format for geographic transactions on the internet.	Spatial/location data
GeoJson	GeoJSON is an open standard format for encoding collections of simple geographical features along with their non-spatial attributes using JavaScript Object Notation.	Spatial/location data

3.1.1 *Format Decision Considerations*

When deciding on what format to release the data, agencies may also consider:

- which open formats are compatible with current system infrastructure
- which formats are widely accepted by developers
- which formats maintain the highest degree of integrity when converted
- if a dataset is proprietary, are there tools available to convert it to an open format e.g. Microsoft Excel file can be saved as a csv
- can the data be released in multiple formats to meet the preferences of the community
- open data maturity.

3.1.2 *Native Format Data*

A dataset may also be made available in its Native (original) format. Although native format data or raw data may not be in an open format, it can provide interpretative benefits and provide users with greater knowledge for development. Agencies can provide native data as an additional file which will be saved alongside the dataset as a resource.

3.1.3 *Multiple Data Formats*

Agencies should consider providing multiple data formats. Additional data formats provide greater scope for developers and the public to re-use the data in format that suites them.

A dataset may also be made available:

- in a widely used proprietary format (e.g. .xls) that encourages re-use; or
- via an existing data tool that allows users to explore, manipulate, and reuse the data.

3.1.4 *Open Data Maturity*

As agencies adopt ‘open data’ practices, the maturity level of their data will develop over time. The World Wide Web Consortium (W3C) has developed a 5 Star Open Data Model that agencies can aspire to.

3 Star Level of Open Data Maturity (Machine-readable and Non-proprietary) is the minimum standard for release of government’s public data for re-use.

It is recognised that agencies need to build capability and maturity in open data. 2 Star Level of Open Data Maturity (machine-readable but proprietary format) may be released in the first release of the data.

The following table provides a summary of the 5 Star Open Data Model:

The 5 Levels	Description of the 5 Levels	Examples
	ON THE WEB, LICENSED FOR REUSE Data is visible, licensed for re-use, but requires considerable effort to re-use	Data is licensed for re-use Tables on a website or data in a PDF document.
	MACHINE READABLE Data is visible, licensed, machine readable and easy to reuse but not necessarily by all	On the web and in proprietary formats.
	NON-PROPRIETORY FORMAT Data is visible, easy to reuse by all and non-proprietary format, which means it is not restricted to specific software.	On the web, machine readable and non-proprietary e.g. CSV file, XML, ATOM; JSON, KML.
	RDF STANDARDS Data is visible, easy to use and described in a standard way. It uses open standards from W3C to identify things so that people can point at your data.	Resource Description Framework (RDF) is a framework for describing resources on the web. It breaks down data into a series of facts.
	LINKED RDF Data is visible, easy to use, described in a standard fashion and its meaning is clarified by being linked to a common definition e.g. linked to other people's data to provide context.	It provides a link to the meaning via a Uniform Resource Identifier (URI). Any other reference to same source means the same meaning can be assumed with confidence.

3.2 FREQUENCY

The Data Authority must make a determination about how often the published datasets will be refreshed.

Consider who is responsible for the datasets ongoing release and maintenance.

This is especially important if data delivery/publishing methods are manual. You may need to establish a process to ensure data is refreshed.

Criteria to consider include:

- how often is the data collected
- the value of data is highest at the point of collection
- will timeliness affect the quality of the datasets
- the resources required for data extract and preparation
- availability classification (refer to availability classification decision)
- will digital services be reliant on the data
- the delivery method for the data.

The frequency for release of the data will be published with the dataset on Data.SA to inform users of the currency of the data.

Frequency options include:

- daily
- weekly
- monthly
- yearly
- quarterly
- As required.

3.3 METHOD OF DELIVERY

The Data Authority and Data Manager will need to determine what the best method to deliver data to the user from an information management system. Where possible the Data Authority should consider data delivery methods that securely automate the release of regular or live open data services direct from our systems.

Open data will be easily discoverable through Data.SA the Government of South Australia Data Directory. Each dataset will have a unique entry on Data.SA that will allow users to search and understand the content of the dataset.

Data.SA currently provides two ways for users to access data:

- hosted on Data.SA (Manual Publishing)
- linked data to an agency portal, data service or automated data delivery.

3.3.1 *Hosted on Data.SA*

Datasets can be stored and hosted on Data.SA. In this instance, files are manually uploaded to the site. Refresh of this data will require a new data file to be manually uploaded to the dataset each time the data is available

.

When to host data on Data.SA:

- data collected periodically where automation costs would exceed benefits (e.g. annual, in frequent or once of data)
- historical data
- system infrastructure does not support automated delivery
- where data protection amendments required cannot be system generated (privacy, secrecy, legislate protected data)

- to release data initially while automation processes are established
- if data is linked and updated automatically then an annual version of the data should be published to preserve the data for historical, research and analysis purposes.

3.3.2 *Linked Data*

Linked data is when data is discoverable on Data.SA and the dataset entry is linked to the source of the data, either an agency portal, data service or through automated data delivery.

When to link data on Data.SA:

- automated data delivery is established
- if an existing data portal or website exists where the agency will maintain the data (either manually or automatically)
- to a data tool that allows users to explore, manipulate, and reuse the data
- if data is of a considerable size (Terabyte+) and a hosting service is used.

3.3.3 *Automated Data Delivery*

It is recommended that agencies consider automating data delivery. Examples of automated data delivery include:

- **Application Programming Interface (API)**
  An API allows your product or service to talk to other products or services. In this way, an API allows you to open up data and functionality to other developers and to other businesses. It is increasingly the way in which agencies and companies exchange data and services, both internally and externally. API's allow developers to build applications that use data. Data that changes rapidly is often delivered through an API. An example of an API is the Australian Tourism Data Warehouse.

  As API's are external to Data.SA agencies need to provide information on how to use the API. Usually instructions and a sample key are provided and published on Data.SA with the link to the API.

  For more information on API's refer to Online Tools and Resources.

- **File Transfer Protocol (FTP)**
  FTP is a standard network protocol used to transfer files. An FTP can securely transfer files from a system to an external location. Agencies may use FTP to deliver data to a FTP address that is linked on Data.SA. The user will still need to download or copy the data when it is delivered. FTP is often used when regular overnight, weekly or monthly data is collected and can be extracted. A web browser can connect to FTP addresses exactly as you would to connect to HTTP addresses. Using a web browser for FTP transfers makes it easy for you to browse large directories, read and retrieve files. A FTP has the potential to scale the automated delivery service to other datasets from the same information system.

- **Web Services**
  A web service is a method of communications between two electronic devices over the World Wide Web and use via standard request-response protocol which allow the exchange of messages in which a requestor sends a request message to a replier system which receives and processes the request, ultimately returning a message in response.

- **Real-Time Data**
  Real-time denotes information that is delivered immediately after collection. There is no delay in the timeliness of the information provided. Real-time data

is often used for navigation or tracking. Agencies are encouraged to research real-time standards and protocols that exist that are appropriate for the data.

View the Government of South Australia's Real time water data and Adelaide metro real time passenger information.

- **Really Simple Syndication. RSS**
  Also called Rich Site Summary or web feeds, RSS is an automated content delivery vehicle in a standard XML file format. RSS feeds benefit users who want to receive timely updates from many sites. As the data is fed out only when data is refreshed it is a popular and simple solution to automation that includes notification.

- **XML Data Extraction**
  Agencies can export a copy of their database into an XML file on a regular basis (nightly or weekly). As the XML file can be quite large it can be compressed into a Zip file. The file is then made available on a web site server. A working example of this method can be viewed at the United States Grants website:

3.3.4 *Selecting the Automated Method*

Before making delivery decisions the following needs to be considered:

- frequency of the data release
- availability classification
- how the data will provide public value
- long term costs savings of investment in automation compared to manual data extract and refresh.

For automated delivery of data, you may also need to consider the following

- additional consideration for webservices, API's, real time data
- developer engagement and support
- relevant standards and protocol
- API basics (refer to Online Tools and Resources.)
- cyber security (consult with your ITSA and/or StateNet Services)
- framework and infrastructure required to maintain the data delivery
- the potential to scale the automated delivery service to other datasets from the same information management system to create efficiencies
- stability of the data structure or information management systems (consider the future of legacy systems or legislative changes that may alter how the content of data delivered)
- emerging technologies
- Automated scripts to remove protected data.

3.3.5 *Additional Consideration for Webservices, API's, Real Time Data*

Reliable automated and sustainable data services (Request-response web services, API's and Real time data) can support economically viable and sustainable digital services that have a great benefit to our community however agencies should also consider:

- testing the concept and benefits
- stakeholder consultation (you must include StateNet Services, your corporate ICT unit and your ITSA)
- determine the approach and design such as protocol, standards (internal and external),hosting design, developer engagement, formats
- consider a single protocol that can be used on several platforms
- Infrastructure and service levels required to support a request-response web services

- establishing gateway limits/ pressuring testing (how many hits on the server to access data)
- developer support such as the framework, documented outputs, technical documentation, notification of updates, forums and terms of service
- risks and mitigation controls (pressure testing, security testing required to protect the reputation of the Government if data services fail.
- system improvement/review of frameworks, protocols, security and pressure test.

3.3.6 *Developer Engagement and Support Eco System*

Where the release of data is through a web service or real-time application early engagement with the community and developers and the establishment of support and communication channels is highly recommended. This may assist you to identify formats, frequency, delivery, and support required to create sustainable value form data release.

The benefits of setting up a developer eco system include:

- you will know who is using your data to develop
- test the concept, improve design and maximise use
- channel to communicate changes, fix problems and find solutions through the community
- developers can act as a first point of call for issues with services and then relay them to you
- emerging standards or technology changes can be discussed.

Simple blogs and online forums can provide low maintenance communication channels that allow two way conversations. These mediums can also be used to supply documentation, frameworks or change notification.

View the Adelaide Metro developer website for a working example of a developer support ecosystem. .

3.3.7 *Licensing your Linked or Automated Data*

If you are publishing data on an agency managed website/tool, ensure that you display a licence logo and statement with the dataset on your website. Note that all dataset entries on Data.SA will be licensed. The licence appears on the bottom left hand side of the dataset page.

Example of CC-BY

This Government of South Australia website is licensed under a Creative Commons Attribution 4.0 License. © Copyright 2014

If you publish several datasets on the same webpage and there are different licence terms required, each dataset will require a licence logo to be displayed in a way that clearly indicates to the user of the dataset what the terms of use are.

Host a Copyright Statement that Supports an Open License

A current open licence Copyright Statement is available on http://www.sa.gov.au/copyright

This statement has been approved for use on South Australian Government websites. You are required to reference your own agency and contact detail in this statement. If you alter any other elements of the statement you will need to seek Crown Solicitor advice.

3.3.8 *Publishing Supporting Resources*

Where ever possible supporting information should be published with your datasets. Agencies often do this through a Zipped file to download.

If you are publishing material to be printed or downloaded as a resource with your dataset you should also ensure this work is licensed appropriately and the material includes a licence logo and statement.

Example provided for CC-BY licensed material.

With the exception of the Government of South Australia brand, logos and any images, this work is licensed under a Creative Commons Attribution 4.0 Licence. To attribute this material, cite the >>Agency>, Government of South Australia, >>title of work<<,>> date the content was sourced<<, >>dataset URL<<

3.4 METADATA

Metadata is the information that defines and describes data. It provides data users with information about the purpose, processes and methods involved in the data collection. There are two types of metadata that the Data Authority should consider in their plans.

3.4.1 *Discovery Metadata*

Discovery metadata is the minimum metadata required for all datasets to make it discoverable on Data.SA. It helps a user decide whether the dataset suits their needs before downloading it.

Discovery metadata is captured when an agency completes the *Data.SA Publishing Content Summary Sheet* (Appendix E).as part of the publishing process (covered in section 9). It includes the name of the dataset, a description, data format, licence, keywords etc.

The fields on the Data.SA Publishing Content Summary are mandatory. Fields include temporal coverage, geospatial coverage, jurisdiction and frequency of release. The metadata fields are defined on the back of the *Data.SA Publishing Content Summary Sheet* (Appendix E).

3.4.2 *Interpretative Metadata*

Agencies should also provide interpretative metadata to accompany their datasets. This type of metadata provides users with more information about the dataset such as its purpose and the methods involved in its collection to help the user understand and interpret the data correctly.

Often industry based metadata standards will be used.

Metadata could include:

- metadata standards requirements
- definition of terms e.g. average weekly earnings
- accuracy e.g. the number of errors
- response rate
- explanatory notes - detailed contextual information, purpose, processes, and methods involved in the data collection

Interpretive metadata needs to be published in an open and machine readable format such as TXT or HTML as separate file to the dataset. Often interpretation metadata will be compressed with the data for download (e.g. zipped). Ensure metadata does not contain information that may breach privacy or security restrictions.

Engage your Data Manager to discuss if interpretation metadata is available and or automated from your information management system.

Consider the *National Statistical Service* Principles for managing metadata View the NSS Principles of metadata and metadata standards that apply to your field of data. The Australian Bureau of Statistics (ABS) offers assistance and advice on metadata. Contact the ABS representative if it is known or contact datasa@sa.gov.au.

3.4.3 *Additional Resources*

Additional resources should also be made available with the data to encourage use or understanding of the data. Any additional files provided will be published alongside the dataset as a resource:

- research reports
- website links which provide context and additional information
- photographs or images of the subject matter e.g. flora.

These resources should be licensed for reuse or copyrights clearly stated to reduce any confusion on how these additional resources can be used.

3.5 DATA DISTRIBUTION SERVICE

Datasets are discoverable on Data.SA, however, your dataset may also be a candidate for a Data Distributor service. A Data Distributor is another data distribution service that collates and distributes data as a single point of truth for a specific type of SA Government data.

If another data distribution service exists, it is recommended that they are engaged to discuss options for data delivery as a **Data Distributor**.

Contact that data distribution service to discuss and confirm the delegation of responsibility as the Data Distributor. They may also take on the role as Data.SA Publisher for the dataset to streamline the process, if they will not then clear process between the parties will need to be defined to ensure that Data.SA is maintained.

These services may also have other processes that will need to be considered, however it is necessary to ensure the open data process has been completed and approved before the Data Distributor releases data on your behalf.

Another Agency may also be your Data Distributor, this often happens as a result of machinery of government changes and infrastructure is maintained by another agency.

3.6 FUNDING OPEN DATA INVESTMENT

The implementation of automated data delivery may require upfront investments depending on the maturity of existing information life cycle management processes at individual agencies. Agencies are encouraged to evaluate current processes and identify implementation opportunities that may result in more efficient use of taxpayer dollars. Where possible open data delivery options should be considered in new ICT enabled projects.

Effective implementation should result in downstream cost savings for the enterprise through:

- information sharing efficiencies
- reduced impact of resources due to freedom of information
- reduction of costs associated with manual publishing open data
- improved policy development based on evidence based data
- scalable automatic delivery service
- maintenance of the data.

Therefore, these potential upfront investments should be considered in the context of their future benefits and be funded appropriately through the agency's capital planning and budget processes.

A detailed business case or cabinet submission may be required for automated data delivery to fund open data investment that requires additional tools or resources.

3.7 COSTS RECOVERY

Data should be available on a non-discriminatory basis to anyone where practicable to encourage its widespread use and to achieve maximum value. Government data will be released free of charge unless:

- the benefits of accessing and using the data are predominantly private in nature rather than creating broader public benefits
- there is a statutory requirement for charges to apply
- Cabinet has approved that charges be applied.

Agencies should refer to the Governments Cost Recovery Guideline (Once approved by Cabinet) for policy direction and guidance on how to develop and review cost recovery arrangements so that they are consistent, transparent and effective.

The Department of Premier and Cabinet in consultation with stakeholder is developing the Governments Cost Recovery Guideline. Until this guide is finalised data that is predominately and exclusively private in nature or has current pricing arrangement in place is out of scope for open data.

APPROACH SUMMARY

Prepare		
☐	Format	Ensure dataset is released in an open format (machine readable and non-proprietary). Multiple formats of data can be released, including native format Identify how data will be transformed into the format for release.
☐	Frequency of Release	Determine frequency of release and what process will be implemented to refresh the data.
☐	Data Delivery	Determine how dataset will be delivered: • hosted on Data.SA (agencies will need to send files to datasa@sa.gov.au for loading) • linked via existing agency portal (enter links to the dataset on the *Data.SA Publishing Content Summary Sheet*) • automated e.g. for an API prepare sample key and any instructions.
☐	Interpretative Metadata & additional resources	Prepare any other material which will help users understand what your dataset is about. This includes interpretive metadata files (csv or txt format), website addresses, reports or images. This will be emailed with the *Data.SA Publishing Content Summary Sheet* to datasa@sa.gov.au.
☐	Data Distributor	Consider if dataset is a candidate for a Data Distributor Service and engage the data distributor in planning the approach for release. Data Distributor may have additional processes. Determine who's responsibility it will be to publish data on Data.SA
☐	Costs	Consider funding or cost recovery implications.
Open Data Process Worksheet		
☐	Record decisions	Record all decisions made about the dataset on the *Open Data Process Worksheet.*