



ENTERPRISE

information services

Open Data Coordinator's Handbook

GUIDANCE FOR MEETING THE OPEN DATA STANDARD
ENTERPRISE INFORMATION SERVICES

CONTENTS

Introduction	2
How to Use this Document	2
The Role of The Data Coordinator	2
Inventorying Data Assets	3
Identifying Publishable Data	3
Inventory Option 1: Identifying Datasets Directly	3
Step 1: Identifying Datasets	3
Inventory Option 2: Identifying Datasets From Business Processes	4
Step 1: Identifying Core Business Processes	4
Step 2: Identifying Datasets from Core Business Processes	6
Inventory Option 3: Identifying Datasets from Source Systems	7
Step 1: Identifying Data Sources (High Level Inventory).....	7
Step 2: Identifying Datasets from Data Sources	8
Finalizing and Submitting your Inventory	8
Prioritizing Data for Release	10
Engaging the public in prioritization	10
Prioritizing Datasets	11
Best Practice: Assess agency’s current open data landscape	13
Planning for Open Data Release	14
Identify open data phase	14
Identify commitments from prioritized datasets.....	14
Write agency open data plan.....	15
Publishing Data	17
Architecting and Preparing for Open Data	17
Obtain Approval.....	17
Curate Fields	17
Publish	18
Contextualize	19
Appendix A: Inventory Field Names and Definitions	20
Appendix B: Commonly Asked Inventory Questions	23

INTRODUCTION

The purpose of this document is to provide guidance to State Agency Data Coordinators and State Agencies as they work to complete the deliverables and requirements identified within the Open Data Standard. The Data Coordinator’s Handbook provides the “how” for agencies as they internal open data programs and inventorying, prioritizing, and publishing open data. This handbook walks through each phase of the open data process (inventory, prioritize, plan, publish) and provides guidance to data coordinators in the completion of each phase of the open data process.

VERSION NOTES

DATE	VERSION	DESCRIPTION OF CHANGES
November 2, 2020	1.0	Soft launch of the Data Coordinator’s Handbook for public review
February 1, 2021	1.0	Final version of Data Coordinator’s Handbook released
September 17, 2021	1.1	Version updated to change “critical datasets” to “core business” datasets to avoid confusion with other uses of the term ‘critical’ in accordance with disaster recovery planning or other agency efforts

HOW TO USE THIS DOCUMENT

This document provides several different approaches to inventorying, prioritizing, planning, and publishing open datasets. Agency data coordinators should use this document as a starting point in completing the deliverables associated with the Open Data Standard and in publishing open datasets and building an open data program within their agency. These approaches are considered guidelines and best practices, and agencies are encouraged to identify and document the procedures that work best for their agency. Each agency should be able to tailor their approach to meet the needs of their agency and to incorporate open data into agency data governance activities.

THE ROLE OF THE DATA COORDINATOR

The data coordinator is a critical and central position, responsible for governance and support of the agency’s open data program and completion of the deliverables within the open data standard. Data coordinators lead the inventory effort, assign work to data stewards, coordinate meetings with subject matter experts to identify and prioritize data, and support the overall governance and management of the agency open data program. Data coordinators act as the single point of contact for all open data activities and coordinate with the Chief Data Officer as needed to support the agency’s open data program needs. The data coordinator should be the approval authority for open datasets being published, with direction as needed from the agency’s data governance body or other executive governance mechanism.

FOR AGENCIES WITH LOTS OF DATA SOURCES OR DATA COLLECTIONS:

Agencies with lots of disparate data sources, over 500 FTEs, or with multiple complex business functions and units should create a network of data stewards to support the data coordinator in completion of their duties. Data stewards act as subject matter experts for specific data sources or divisions and can provide overall governance and support to the data coordinator as they inventory, prioritize, and plan to publish open data. Data stewards should be knowledgeable in the sources, collection methods, and limitations of specific datasets or data systems and collaborate with the data coordinator

in curating and preparing data for publication. Data stewards are highly beneficial and allow the agency to save time and share resources more effectively in the management of open data.

INVENTORYING DATA ASSETS

The data inventory is the first deliverable associated with the open data standard. Completion of the inventory is a requirement, but there are many approaches an agency can take in how they identify their information assets. Identified below are a few different approaches an agency may select, as well as guidance on which methods may be more relevant for specific agency sizes. While the approach is flexible, agencies should use the data inventory spreadsheet consistently and provide as much information as possible about their data assets when submitting the completed inventory sheet.

Agencies should focus on identifying and inventorying core business and enterprise data assets initially, rather than working to identify every single piece of data that exists within the organization. Core business or enterprise datasets are:

- Data that are used in the core business functions of the agency, such as administering programs, providing benefits or services
- Data that is or has been used in decision making, documents performance, transactions, or assets
- Data that is relied upon by multiple individuals or agencies in the creation, management, or oversight of a program, service, or project
- Data that is frequently requested through the public records process or has a high level of public interest and engagement

IDENTIFYING PUBLISHABLE DATA

The inventory process requires that agencies identify “publishable” datasets within their inventory. Publishable data is defined as any data and datasets collected by a state agency that the agency cannot withhold access to due to its sensitivity, competitive advantage, or other limitations. For the purposes of the inventory, the agency should indicate where the data falls under the following definitions of publishable:

- **Data can be published as is:** the data can be published with no redactions or removal of fields
- **Data can be published with some redactions:** the data may be published but would require redactions, de-identification, statistical noise, or aggregation to make it public
- **Data is not publishable:** the data cannot be published in any format.

The existence of a few sensitive fields does not mean the dataset may be withheld in its entirety from publication. If the data is not publishable in its raw format, the agency must identify what processing can occur to make the data publishable. As an example, if the raw data contains personally identifiable information, could the agency withhold or redact the PII elements and publish the dataset without breaching privacy or breaking protocols?

INVENTORY OPTION 1: IDENTIFYING DATASETS DIRECTLY

Best suited for: small agencies, agencies with only a few datasets, or agencies with a single business subject matter expert who can identify all source systems and data assets used by the agency with minimal need to interview stakeholders.

STEP 1: IDENTIFYING DATASETS

Using the inventory spreadsheet provided, list all of the datasets controlled, regularly created, or maintained by, or on behalf of your agency. These datasets should encompass all of the different functions and systems your agency may use, or

if you are responsible for aggregating data from other government entities (e.g., receiving data submissions from county, local, or special service district governments); encompass the various data collections your agency undertakes.

QUESTIONS TO ASSIST YOU IN IDENTIFYING DATASETS:

- Where does your agency capture information? Are there forms, systems, or places where clients or end-users enter information?
- What information is your agency publishing on its website in the form of reports, datasets, or statistical analyses? Where did the data for those resources come from?
- Are there any central Access databases, Excel spreadsheets, or other sources where multiple people are inputting data?
- Which systems have been identified in either a Disaster Recovery or Continuity of Operations Plan as being vital to your organization? Which datasets are typically reported on from this specific data source? Applications submitted, clients served, etc.
- Which information is used to populate your agency's performance measures or annual reports?
- Which data does your agency collect from or aggregate from cities, counties, special districts, or other regional governments?
- Are there datasets your agency is regularly required to report to the federal government?
- Are there datasets that are regularly requested by outside entities or constituents (including public records requests)?
- Outside of your data sources, are there other datasets or resources your Agency is already publishing online in the form of reports, analyses, or dashboards?

For each identified dataset above, completed the information on the inventory spreadsheet. A list of all fields and their formats is listed in Appendix A.

INVENTORY OPTION 2: IDENTIFYING DATASETS FROM BUSINESS PROCESSES

Best suited for: larger agencies or agencies with lots of program areas, agencies who would prefer to look at business processes rather than source systems (see Inventory Option 3)

In this approach, the Data Coordinator will review current program documents, strategic plans, and other materials to identify the core business functions or processes of the agency. These core business functions can then be used to determine which datasets are used, maintained, managed, or created as part of those business functions.

STEP 1: IDENTIFYING CORE BUSINESS PROCESSES

Step 1 of this inventory process will provide you with a high-level inventory of the business functions and programs supported by your agency. The inventory spreadsheet contains a template document that you can use to note business functions, activities, and information sources created as part of those functions.

To help identify the business functions of your agency, you may wish to start by reviewing reports and documentation to help identify the high-level work your agency performs, such as:

- Organizational Charts
- Strategic plans, business plans, or other planning documents
- Agency policies and procedures
- Legislation or Statute governing your organization

Once you have a high-level understanding of the different divisions, business units and program areas of your agency, conduct stakeholder interviews within each business unit of the organization and identify the core business processes for each program.

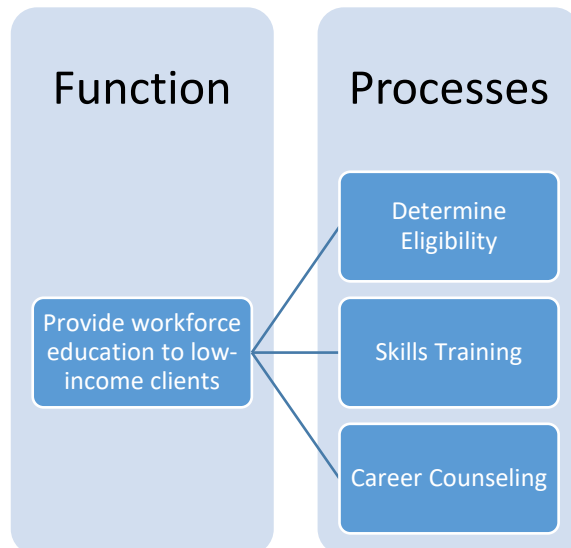
Stakeholder interviews will allow you to work collaboratively with the business unit to identify core business processes and functions and the datasets that support those functions.

HELPFUL QUESTIONS FOR STAKEHOLDER INTERVIEWS:

- What functions and activities is your area responsible for?
- Which functions are considered 'mission critical' in the event of a disaster or situation that would require you to provide only minimum service levels?
- Which shared functions between your agency and another agency require the sharing or transmission of data from one program, agency or partner to another?
- Which constituent populations does your area serve (e.g. children, parents, clients)?
- What other populations are served by your program (e.g. businesses, hospitals, animal populations)?
- What types of services does your program provide?
- Which systems do you use most frequently to provide or manage services?
- Are there rules or legislation that govern how you conduct business?
- Are there external or internal dependencies on specific business functions?

After conducting interviews, you should complete an inventory of business functions and processes or activities. Two different example inventories are provided below, to show some options available in completing your inventory and in the level of detail you wish to capture. There is a sample function inventory available in the inventory spreadsheet provided by the Chief Data Office.

EXAMPLE: FUNCTION-PROCESS INVENTORY



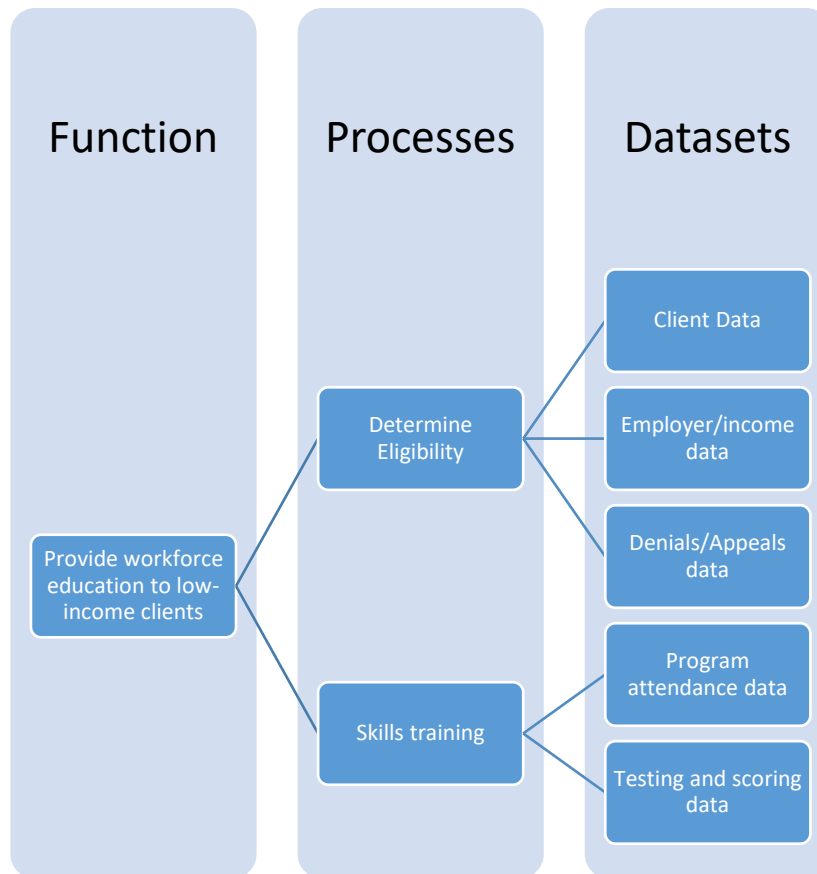
EXAMPLE: ALTERNATE FUNCTION-PROCESS INVENTORY

Function	Process	Sub-process
Provide workforce training to low-income clients	Determine eligibility for services	Income verification
		Acceptance/Denial/Appeal
	Skills Training	Tracking attendance
		Testing and learning assessments
	Career Counseling	Identifying career affinity
		Counselor 1:1s
		Rotational Placement

STEP 2: IDENTIFYING DATASETS FROM CORE BUSINESS PROCESSES

In reviewing the business processes you have identified, brainstorm the datasets that go into that business process. For example, in “determining eligibility for services” you may have a client dataset, an income dataset, or a denial/appeals dataset.

EXAMPLE: IDENTIFYING DATASETS FROM FUNCTIONS



QUESTIONS TO ASSIST YOU IN IDENTIFYING DATASETS:

- What types of information are collected as a result of this process?
- Which systems are used in the management of this process?
- What kinds of files, documents, forms, or data are we required to maintain as part of this process?
- What metrics do we report on as part of this function or business unit?
- What information do constituents, the media, or others regularly request access to, either through the public records process or other means related to this process?

For each identified dataset, completed the information on the inventory spreadsheet. A list of all fields and their formats is listed in Appendix A.

INVENTORY OPTION 3: IDENTIFYING DATASETS FROM SOURCE SYSTEMS

Best suited for: larger agencies or agencies with lots of program areas, agencies who have already completed an application inventory or have a listing of data systems/core applications

STEP 1: IDENTIFYING DATA SOURCES (HIGH LEVEL INVENTORY)

Step 1 of this inventory process will help you map out where potential datasets reside within your Agency by producing a high-level inventory of data and information systems.

Potential Data Sources in your Agency may be databases, applications, or other “systems of record” that are considered vital to the continued functioning of your organization. These systems could be cloud-based, on premise, in shared drives, or other sources.

QUESTIONS TO ASSIST YOU IN IDENTIFYING SOURCE SYSTEMS:

- Which IT or Information Systems does your agency use on a regular basis?
- Which applications or solutions are used to capture information (e.g. online forms, applications, or places where clients may enter information)?
- What information is your agency publishing on its website in the form of reports, datasets, or statistical analyses? Where did the data for those resources come from?
- Are there any central Access databases, Excel spreadsheets, or other sources where multiple people are inputting data?
- Which systems have been identified in either a Disaster Recovery or Continuity of Operations Plan as being vital to your organization?
- Are you currently housing any systems at the State Data Center?

For each data source on the high-level inventory, document the following fields on the “Optional Data Source Inventory” tab of your inventory spreadsheet:

Field Name	Definition
Division	The division or sub-agency responsible for managing/governing the system, or the primary user of the data source
Data Source Name	The name of the system, database, or software (e.g. “Online Licensing System,” “E-Permitting Application”)

Description	A description of the system’s purpose and primary types of information held. Example: “The E-Permitting System is used for any Oregon Business seeking a permit to build or expand a building. It contains applications, petitions for review, appeals, notices of denial/approval, and addresses”
Technical Details	Notes about the software version or whether the system is on premise or a cloud based system
Key Contact(s)/SME(s)	Any business owner with in-depth knowledge about the types of data within the system

STEP 2: IDENTIFYING DATASETS FROM DATA SOURCES

Use your list of data sources to identify potential datasets. In some instances, where your data source is a spreadsheet or single table, there may only be one associated dataset. For larger applications or relational databases, there will be multiple potential datasets residing within the system. It may help to collaborate with the subject matter expert or business owner with knowledge of the system to identify possible datasets.

Additionally, your agency may be producing or aggregating datasets outside of the data sources identified in Step 1. Use some of the guidance below to help identify potential datasets residing throughout the organization.

QUESTIONS TO ASSIST YOU IN IDENTIFYING DATASETS:

- Which datasets are typically reported on from this specific data source? Applications submitted, clients served, etc.
- Which information is used to populate your agency’s performance measures or annual reports?
- Which data does your agency collect from or aggregate from cities, counties, special districts, or other regional governments?
- Are there datasets your agency is regularly required to report to the federal government?
- Are there datasets that are regularly requested by outside entities or constituents (including public records requests)?
- Outside of your data sources, are there other datasets or resources your Agency is already publishing online in the form of reports, analyses, or dashboards?

For each identified dataset, completed the information on the inventory spreadsheet. A list of all fields and their formats is listed in Appendix A.

FINALIZING AND SUBMITTING YOUR INVENTORY

Once you have identified your agency’s datasets, coordinate with a few agency subject matter experts to review the inventory and ensure that no core datasets have been overlooked. For an initial inventory, the agency should do their best to identify core business datasets that are central to the work of the agency. As new business processes or systems are added, you can update your inventory in the future.

If all stakeholders/subject matter experts have reviewed and signed off in the inventory as being as complete as possible, and once the agency has engaged the prioritization process identified below, submit your data inventory to kathryn.helms@oregon.gov with the subject line: [Agency Name]: Data Inventory. Agencies should prioritize datasets before submitting their finalized inventory.

Agencies will receive email confirmation that their inventory has been received, along with a report indicating any suggested changes that need to be made to the data inventory (e.g., correct missing required fields, update invalid inputs) before it is added to the enterprise data inventory and published on data.oregon.gov.

PRIORITIZING DATA FOR RELEASE

After completing the Agency Data Inventory, agencies begin reviewing data assets within the inventory to prioritize these datasets for publication. Datasets should be prioritized based upon their evaluated value as an open dataset, in addition to the level of effort involved in curating the datasets for publication. Additionally, data coordinators should evaluate currently available reports and data made publicly accessible by the agency and evaluate these offerings against the 5-star model of open data to identify areas for the agency to mature its data publication efforts.

The following steps exist to provide agencies multiple avenues to assess their current open data offerings, identify datasets that may be of value for release, and begin planning for publication of open datasets. Agencies may undertake their own internal prioritization processes or utilize other means to prioritize their open data release as they see fit. Agencies should document replicable, sustainable procedures for data prioritization so that the agency's decision-making around open data is clearly documented and future prioritization efforts can follow established agency procedures.

For the purposes of the inventory, agencies identify data as belonging in one of four priority categories:

- High – this dataset is a high priority for release as an open dataset
- Medium – this dataset is a medium priority for release as an open dataset
- Low – this dataset is a low priority for release as an open dataset
- No priority – this data is either already released or cannot be released due to privacy concerns and therefore does not require prioritization

ENGAGING THE PUBLIC IN PRIORITIZATION

Open data allows for creation of a greater feedback loop between constituents and government, and state agencies should leverage this opportunity to help prioritize future open data releases. When examining open data commitments, agencies should consider the following opportunities for public engagement now and in future years:

- **Examine suggested datasets.** Constituents are able to suggest or request datasets on the [state's data portal](#), and is recommended that agencies review these current suggestions to see if there are datasets, they can produce in response to these requests.
- **Review common records requests.** Agencies can work with their communications staff to determine which datasets constituents frequently request and prioritize these datasets for release. Examination of public records requests should be an ongoing process and the agency data coordinator should stay abreast of current popular topics or requests by coordinating with communications or public records staff.
- **Determine what is already popular.** If your agency already publishes reports or data, assess which assets are most popular or frequently downloaded and work to release complementary datasets, or enhance current offerings.

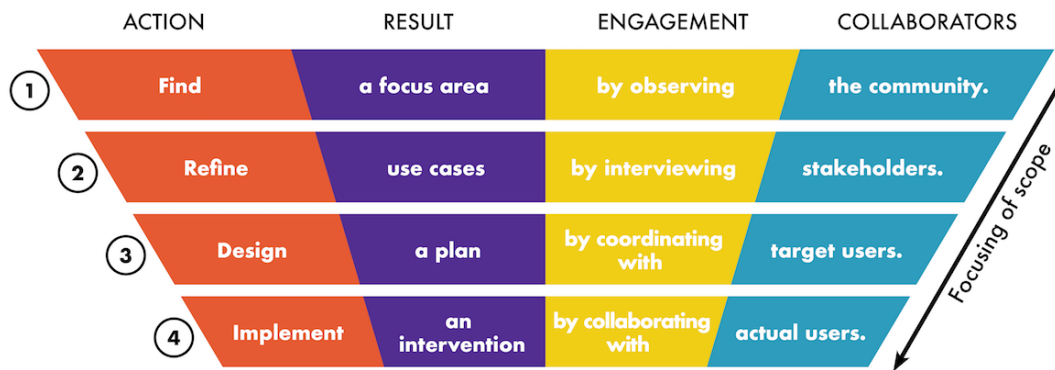
TACTICAL DATA ENGAGEMENT: A FRAMEWORK FOR APPROACHING PRIORITIZATION

The Sunlight Foundation's [Tactical Data Engagement](#)¹ playbook identifies a four-step process to prioritizing and releasing open data with constituent and stakeholder feedback in mind. This process utilizes both internal prioritization exercises and

¹ "Tactical Data Engagement — Roadmap to Informed Communities."

public outreach to assist agencies in identifying datasets that would be of use to constituents and meet the needs of the agency as an open dataset. Agency data coordinators should consider the resources available through the Sunlight Foundation as a starting point for building internal open data and creating an open data culture within their agency.

Tactical Data Engagement: The 4-Step Process



1 The Tactical Data Engagement Process from the Sunlight Foundation

As agencies build and maintain their open data programs, prioritization will continue to be an iterative process and relies upon building relationships with stakeholders and constituents to help identify open data needs.

PRIORITIZING DATASETS

Data coordinators should work closely with agency leadership, stakeholders, communications staff, and data subject matter experts to prioritize datasets for release. The prioritization process is designed to get the agency to think critically about which high-value datasets are currently maintained within the agency, their value as an open dataset, and their overall readiness for publication. Data coordinators may benefit from a series of workshops or conversations with data users and staff who handle public records requests to get a strong picture of datasets that would be beneficial as open datasets.

DATA VALUE

Data value is the relative value or impact of publishing this data as an open dataset. Data value can be determined by asking several questions of the data:

HAVE AN OPEN DATASET IN MIND?

Vet your idea with the [Sunlight Foundation's Open Data Quiz](#)

- Do constituents, government entities, or other stakeholders regularly request this dataset?
- Would the release of this data contribute to government transparency and accountability in a meaningful way?
- Is the information contained in this dataset of high value to the public?
- Does this dataset have the potential to support inter-agency data initiatives or larger scale analysis projects?
- Does this dataset represent a core service or mission of the agency? Is this dataset used in strategic planning or decision-making for the agency?
- Is this dataset part of an already published report or other documentation that is frequently used and accessed?
- Could the release of this dataset present an opportunity for a new application, technology, or tool that could provide some benefit to the agency or the constituents it serves?

DATA READINESS

Data readiness refers to the current quality and completeness of the data to be published. Data readiness can be determined by asking several questions of the data:

- Is the data consumer ready? If the dataset is not consumer ready, what level of effort would be required to process, fill-in, or otherwise prepare the data for publication?
- Are calculated data values well documented, and is there a supporting data dictionary for this data?
- If there is no data dictionary, what level of effort would be required to create a data dictionary?
- Are field definitions consistently applied throughout this dataset?
- Would the process of publishing this data be cost-prohibitive?
- Can the publication of this dataset be automated through creation of a database view, or using a product like [Socrata Gateway](#) to synchronize with data.oregon.gov automatically?
- Are there quality or integrity issues with the data? What level of effort would be required to remediate these issues?

FOR DATASETS WITH PRIVACY CONSIDERATIONS

For datasets on the inventory with private information, such as personally identifiable information (PII) or private health information (PHI), agencies should consider the following additional questions when examining both the value and readiness of a dataset for publication. Agencies should consider creating either a privacy impact assessment or risk analysis process for evaluating datasets with PII for publication/redaction, and establish procedures to

- Is there personally identifiable information (PII) or private data within the dataset? If so, consider completing a Privacy Impact Assessment to evaluate the value of releasing the dataset and the level of redaction required. Further privacy information can be found in “Open Data Privacy: Guidance for Agencies”
- Is this dataset publishable in its raw format or will there be aggregation, anonymization, or other obfuscation actions required? Is the dataset usable or informative in a redacted format?

PRIORITIZING DATA BASED UPON VALUE AND READINESS

Using the two rankings of value and readiness, agencies can identify the overall priority of a dataset for publication based upon the matrix below. Data that is of high quality and is highly ready for publication should be the highest priority for publication as an open dataset. Some of these high priority datasets may be currently provided data offerings that the agency has identified as needing maturation or publication in a more open, usable format, per the 5-star open data model below.

OPEN DATA PRIORITIZATION MATRIX				
Value of Data Publication	High	Mid Priority	High Priority	Highest Priority
	Medium	Low Priority	Mid Priority	High Priority
	Low	Lowest Priority	Low Priority	Mid Priority

		Low	Medium	High
Readiness of Data for Publication				

BEST PRACTICE: ASSESS AGENCY'S CURRENT OPEN DATA LANDSCAPE

Maturation or improvement of current data offerings is a useful avenue for publishing open data. During the inventory process, agencies will have identified datasets that are already made publicly available as open datasets, dashboards, reports, or other formats. By reviewing the datasets that are already made publicly available, agencies can begin to assess their current open data landscape.

Agency data coordinators should look closely at datasets or resources that are currently published online and evaluate whether this data is truly open, or if it remains locked into PDF reports and other static reporting mechanisms. Open data commitments utilizing already publicly accessible data provide an easy starting point in automating and developing

mechanisms to support open data internally, and provide an operational benefit by automating already extant work processes.

WHAT ABOUT DASHBOARDS?

Dashboards are a tool to make data more interactive and engaging for end-users, however, they are not open data. Agencies publishing dashboards should explore publishing the underlying data in their dashboards as a companion dataset.


EXAMINE DATA ALREADY BEING PUBLISHED: THE 5-STAR MODEL OF OPEN DATA

The Five Star Model for Open Data² is a tool to determine the current level of openness for datasets or information an agency is already publishing online. By evaluating currently available data or tables on an agency website or data portal, data coordinators can identify data that would benefit from moving to a five-star rating, while maturing the agency's

overall open data maturity.

Number of Stars	Definition	Example
★	Data is available on the internet in some form under an open license.	A PDF report with statistical information or agency information. An embedded image of a table on a webpage
★★	Data is available as machine-readable or structured data.	A downloadable excel file.
★★★	Data is available as machine readable or structured data in an open/non-proprietary format.	A downloadable .csv file or other open format.
★★★★★	All of the above, plus data is available for download via an API endpoint	A documented API exists for accessing the data. Socrata (data.oregon.gov) and ArcGIS Online enable this natively through their applications.

² The 5-star Deployment Scheme for Open Data, adapted from the model created by Tim Berners-Lee. <https://5stardata.info/en/>

	All of the above, plus data is appropriately described and contextualized	Data is published with a companion data dictionary and appropriate levels of metadata to make it easily discoverable and understandable.
---	---	--

In reviewing their current published data offerings, agencies should assess whether some of the data or reports they currently publish could benefit from being improved or moved closer to a five-star open data rating. As an example, an agency who publishes a PDF report with statistical information that is popular amongst constituents may benefit from turning the data within the report into a companion open dataset on data.oregon.gov.

PLANNING FOR OPEN DATA RELEASE

The deliverable for this step is creation of the Agency Open Data Plan, which should answer the question “What will my agency commit to regarding open data through the next biennium, and what will we plan for in future biennia?” Agencies may commit to publishing new data assets, maturing current data offerings, or enter into a planning phase to prepare data assets for publication.

IDENTIFY OPEN DATA PHASE

Agency data coordinators and stakeholders should enter into a strategic conversation based upon their high priority datasets and the time and resourcing commitments required to publish these datasets as open data. Agencies will also identify the appropriate phase they are in concerning open data from the following options:

- **Planning:** if the agency does not anticipate publishing any datasets due to the preparation and resourcing required to publish data during the current biennium, the agency should still be engaging in planning and preparation efforts to publish datasets in the future. Agencies cannot be in the planning phase for more than one consecutive biennium. The agency open data plan should reflect which datasets the agency is planning to publish in the future.
- **Publishing:** If the agency intends to publish open data this biennium, the open data plan should reflect which datasets the agency intends to publish and any additional detail on data categories or types of data the agency may intend to publish.
- **Maintenance:** If all of the publishable data identified in the agency’s inventory has been published and cataloged appropriately according to the agency’s data inventory and prioritization outcomes, the agency is in a maintenance phase.

For agencies in the planning phase, they should identify either candidate datasets or explicit planning activities the agency will undertake over the coming years to publish open data in future years. Agencies who enter into a planning phase may also consider maturing one or two currently published reports or data offerings as a pilot for open data in advance of planning and publishing for more complex datasets.

IDENTIFY COMMITMENTS FROM PRIORITIZED DATASETS

If the agency is able to release datasets during this biennium as part of a publishing phase, how many datasets can the agency realistically publish during this timeframe? Are there current projects or initiatives underway within the agency that could further support or enable the publication of open datasets? If the agency is not yet ready to publish open datasets due to outside constraints, what planning activities will the agency undertake this biennium to be ready to publish in the future?

Open data commitments should include one of the following:

- Improving an existing open dataset, which includes one or more of the following actions:
 - Changing the data schema (e.g., adding and/or removing columns)
 - Increasing granularity or making a dataset more specific (i.e., aggregated less or not at all)
 - Converting a manual dataset to an automated process
 - Increase scope of data in a dataset (e.g., a dataset covering a single permit type may be expanded to cover multiple permit types.)
- Publishing a new dataset, which includes one or more of the following actions:
 - Planning for data publication, and/or researching the feasibility of extracting data from agency information systems
 - Architecting a process for extracting data and publishing to data.oregon.gov
 - Drafting metadata and a data dictionary
 - Publishing the dataset on data.oregon.gov
 - Automating dataset updates/publication
- Cataloging an external dataset on data.oregon.gov
- Creating a new dashboard/visualization/Socrata Story to accompany an open dataset or externally published data

WRITE AGENCY OPEN DATA PLAN

The Agency Open Data plan should be 2-3 paragraphs, including a brief description of the agency’s core mission and functions, identification of which phase the agency is in as listed above, and a description of the datasets the agency plans to release during the biennium. If the agency is in a “planning” phase for open data, there should be a description of the planning and preparation work the agency will undertake in the current biennia and which datasets the agency is planning to publish in the next biennium. Agency open data plans must be submitted to the Chief Data Officer on a biennial basis. The first deadline for the Agency Open Data Plan is June 1, 2022, though agencies can submit their plans as soon as they are ready.

SAMPLE OPEN DATA PLAN 1: ENTERPRISE INFORMATION SERVICES

Enterprise Information Services (EIS) provides statewide IT leadership in the areas of IT governance, cyber security, data and information governance, shared IT services, and data center services. Enterprise Information Services works to ensure user-friendly, reliable, and secure state technology systems that serve Oregonians. EIS provides services that span a number of agency and technology needs, from an enterprise telephony system and digital services, data center operations and maintenance, to IT project management and cyber security assessments. The datasets captured within the EIS inventory encapsulate the data created as each program area carries out their daily work.

EIS has inventoried 64 datasets and identified datasets related to IT Portfolio Management and Oversight as the most valuable resources to publish this biennium. These datasets focus on total IT investment and portfolio health for all state agencies who are required to work with EIS when implementing new technology solutions. Publication of these datasets will begin at the beginning of the fiscal year and should encompass 1-2 datasets on data.oregon.gov as well as a project dashboard for individuals to interact with the data that has been published.

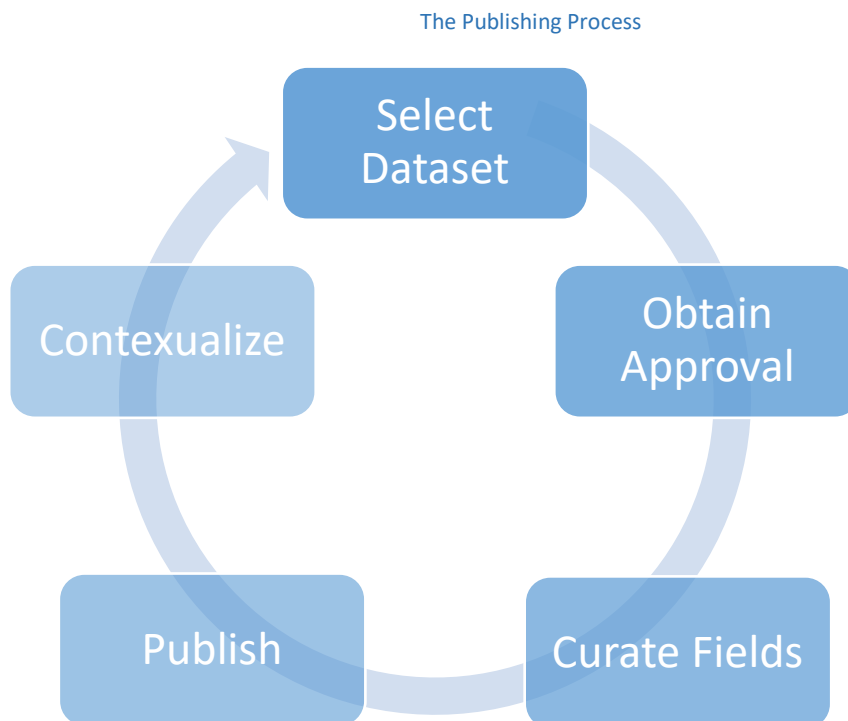
SAMPLE OPEN DATA PLAN 2: DEPARTMENT OF CONSUMER AND BUSINESS SERVICES

The Oregon Department of Consumer and Business Services (DCBS) works to protect and serve Oregon's consumers and workers while supporting a positive business climate. As Oregon's largest consumer protection and business regulatory agency, DCBS is a resource to consumers and businesses in areas involving: finance (insurance, investments, Oregon-chartered banks and credit unions, consumer finance companies, mortgage lenders, pawnbrokers, payday lenders), workplace health and safety, and building safety. The datasets identified in DCBS' inventory embody the data created as each division and program area carries out their regulatory activities.

DCBS has completed an initial inventory and identified data related to Oregon OSHA's enforcement as the most ready and valued dataset to publish this biennium. This dataset covers Oregon OSHA's enforcement activities, including inspections, citations, violations, penalties, and accident investigations. Publication of this dataset will begin at the beginning of the fiscal year and should encompass 2-4 datasets on data.oregon.gov.

ARCHITECTING AND PREPARING FOR OPEN DATA

Once an agency has identified their open data commitments, they should begin planning and preparing for the publication of their chosen datasets. The process of publishing open data involves time spent obtaining any approvals needed from data stewards or data subject matter experts, curating the appropriate fields to appear in the dataset, designing a publication process or method, and adding appropriate context to the data.



OBTAIN APPROVAL

The Open Data Plan should be authored with feedback from data providers and subject matter experts, and Data Coordinators should make sure that they have agency leadership and data subject matter expert approval to publish the prioritized datasets. Once a dataset has been selected, the Data Coordinator should convene a small working group of subject matter experts and any technical administrators of the dataset to document any limitations on the dataset, the technical architecture of the source system, and any other considerations the agency needs to address in the process of publishing the identified dataset. Agencies are encouraged to develop a governance and approval workflow for datasets, where subject matter experts review the dataset before publication and sign off on the dataset and additional documentation, such as data dictionaries and metadata.

CURATE FIELDS

After selecting a dataset to publish, the agency Data Coordinator should work with subject matter experts and any individuals with authority over the dataset to determine which fields are most appropriate and relevant to publish, in

addition to documenting any privacy or security limitations that may require the data be aggregated, anonymized, or otherwise transformed to become a publishable dataset. The agency should create a listing of all fields to include within the dataset, and begin structuring column order, column names, and any human-readable or display names that will be added to the open dataset.

Questions to consider as agencies determine the field listing and data elements to include in the open dataset are:

- Does this data have a unique identifier that can be published to enable easier updates?
- Are there specific fields, dimensions, and calculated fields that need to be included for this dataset to be meaningful?
- Is there Personally Identifiable Information (PII) or Protected Health Information (PHI) in this dataset? If so, should the agency perform a privacy assessment to determine how best to release this data?
- Are there standard codes or unique fields that would make this dataset more interoperable (e.g. utilizing common standards such as FIPS codes, NAICS codes, etc.)?
- Are there combined or concatenated fields that would benefit from being split into multiple fields, such as an address that could be split into Street Address, City, State, and Zip code fields?
- What number formatting, if any, needs to be applied to this data after it is exported from a source system?
- Can this dataset be turned into a view within a database or similar data table to allow for automated export or updating of the dataset once it is published?

As the agency builds their dataset, remember that the audience for open data is broad and may include individuals and organizations who do not yet have subject matter expertise on the dataset or even the agency's lines of business. Err on the side of choosing plain-language column headings, documenting calculated fields, and providing as much context as possible in curating a dataset.

PUBLISH

DATA.OREGON.GOV

The State Open Data Portal, data.oregon.gov, is a vendor-provided solution called Socrata Data Management Platform. Socrata automatically makes each published dataset available to download in multiple formats, accessible through an Application Programming Interface (API), and contains some charting/visualization capabilities within the application itself, allowing users to filter, sort, roll-up, download, and visualize data within the platform. The Socrata Data Management Platform allows for multiple options in developing a workflow for open data, automating data publication to the portal, and provides mechanisms for both data consumers and publishers to interact with datasets made available on the portal through APIs

PUBLISHING DATASETS OUTSIDE OF DATA.OREGON.GOV

Agencies are expected to utilize data.oregon.gov wherever possible in publishing open data, although there may be use cases where a different platform, such as ArcGIS Online, is a better option for a specific dataset. Agencies may publish datasets outside of data.oregon.gov, as long as the mechanism for publishing allows for the same level of data interaction (data must be accessible in a machine readable, open format, downloadable through an API endpoint or similar) and these datasets must be federated/cataloged in data.oregon.gov to maintain accessibility. Agencies should work directly with the Chief Data Officer if they have external portals they plan to use in lieu of data.oregon.gov.

The state also leverages ESRI's ArcGIS Online tool for the publication of geospatial data, and agencies may continue using ArcGIS Online in the publication of geospatial datasets. Datasets that have a location component such as addresses, zip

codes, city blocks, but are primarily structured non-spatial datasets may be better served by utilizing data.oregon.gov. Assets published in ArcGIS online are federated (linked) between the ArcGIS platform and Socrata, so agencies do not need to put additional effort into cataloging these items. For additional information on utilizing ArcGIS Online or other geospatial data platforms to meet the requirements of the Open Data Standard, refer to “Guidance for Geospatial Data Publishers” at the end of this section.

CONTEXTUALIZE

Adding context to data through metadata and data dictionaries increases the findability and accessible of datasets published on data.oregon.gov. Agencies should plan carefully to include relevant metadata and data dictionaries for all datasets they publish, and expect to put some effort into adding useful context to datasets they publish.

METADATA

Metadata, or “data about data,” provides valuable contextual information about a dataset in addition to improving the dataset’s discoverability within the data.oregon.gov catalog.

DATA DICTIONARY

Agencies are expected to provide a data dictionary for each dataset published to data.oregon.gov. The data dictionary provides a detailed listing of all fields in the dataset, a description of the field, and information about input values. Agencies may also wish to provide additional resources, such as schema diagrams or other artifacts, if the agency is publishing several datasets from the same database or multiple tables from a single relational data source. Including data schemas and diagrams allows users to recreate any data relationships that exist within the source system and more easily interact with and analyze open data.

APPENDIX A: INVENTORY FIELD NAMES AND DEFINITIONS

Column Name	Description	Required	Input
<i>Department/Division</i>	Name of the business unit who is responsible for managing this data	Yes	Free text
<i>Dataset name</i>	Brief descriptive name for the dataset	Yes	Free text
<i>Brief description of data</i>	Include a brief description of the dataset. What is the purpose? What is it used for? Include key data fields if possible.	Yes	Free text
<i>Subject Matter Expert - Position Title</i>	Position title for the individual who knows the most about the data, its origins, and can answer questions about it or route questions to the appropriate parties.	Yes	Free text
<i>Subject Matter Expert - Name</i>	Name of the Subject Matter Expert	Yes	Free text
<i>Subject Matter Expert - Email</i>	Email of the Subject Matter Expert	Yes	Email
<i>Geographic boundaries</i>	If this data contains geographic information, what are the bounding boxes for this data?	Only if applicable	Free text
<i>Geographic granularity</i>	What is the lowest level of geographic granularity in the data? For example, if the data is collected by address, it would be Street Address.	Only if applicable	Free text
<i>Frequency of data change</i>	At what rate does the information in the dataset change? E.g., Hourly, Daily, Monthly, Annually	Yes	
<i>Existing publication</i>	Is this data already made available for public consumption through an existing report, API endpoint, portal, etc.?	Yes	Yes or No
<i>Type of existing publication</i>	If you answered yes to the previous question, please include a link to the publication or a description of the type of publication	If "Yes" to previous question	Hyperlink or Free text
<i>Data Classification</i>	How would you classify this data? Level 1, "Published" - Information that is not protected from disclosure that will not jeopardize the privacy or security of agency employees, clients or partners. Level 2, "Limited" - Information that may not be protected from disclosure, but may jeopardize the privacy or security of agency employees, clients, or partners if made readily available. Examples: risk assessments, internal audits. Level 3, "Restricted" - Information intended for limited business use that may be exempt from public disclosure due to risks posed to privacy and security associated with release of this information. Examples: network diagrams, personally identifiable information, information exempted from public	Yes	Level 1, Level 2, Level 3 or Level 4

	<p>records disclosure.</p> <p>Level 4, "Critical" - Information that is considered highly sensitive and is intended for use by specified individuals only. Disclosure or release of this information poses a series threat and could cause major damage or injury. Examples: Records of victims of domestic violence/stalking, IRS records/data</p>		
<i>Data contains PII/PHI</i>	Does this data contain either Personally Identifiable Information or Personal Health Information?	Only if applicable	PII, PHI, PII/PHI, No
<i>Data Restrictions or Regulations</i>	Are there other restrictions or regulations that would limit the release of this data or require redaction of some fields? E.g. FERPA, HIPAA, Public Records Exemptions, statutory or contractual limitations.	Only if applicable	Free text
<i>Data is/is not Publishable</i>	<p>Is this data publishable?</p> <p>Data can be published as is: the data can be published with no redactions or removal of fields</p> <p>Data can be published with some redactions: the data may be published but would require redactions, de-identification, statistical noise, or aggregation to make it public</p> <p>Data is not publishable: the data cannot be published in any format.</p>	Yes	As is, Redaction required, Not publishable
<i>Data Prioritization Fields: complete after initial inventory</i>			
<i>Data Value</i>	<p>What is your sense of the relative value in publishing this data?</p> <p>High - existing and ongoing requests for this data; this data addresses pressing information needs or pain points (for either State government or other entities); or we have heard compelling examples of how this data could be used;</p> <p>Medium - This data may be useful for other state or government agencies or for other constituents; we occasionally receive requests for this information; or we have heard some examples for how this data could be used;</p> <p>Low - This data has unclear value for either the public or other state or government entities; we have never received requests for this data; or we have never heard a use case for this data</p>	Yes	High, Medium, or Low
<i>Data Value Comments</i>	Are there details about this data that make it a particularly high priority or low priority to publish? Provide comments if needed. Example: "Data is	No	Free text

Data Readiness for Publication

requested frequently in public records requests" or "data is low quality and does not present much public value if posted"

What is your sense of the readiness of this data to be made publicly available?
High - this data is of a high quality, requires little to no de-identification, intervention or transformation to make it suitable for external users
Medium - This data may have a few quality issues, technological challenges/limitations, or other elements that would require work on behalf of the agency to clean, prepare, and ready it for publication
Low - this data is of a low quality/usability, is technically challenging to access and export, or has other limitations that would require extensive work on behalf of the agency to clean, prepare, and ready it for publication

Yes

High, Medium, or Low

Data Readiness Comments

Are there details about this data that make it either ready for publication or unsuitable for publication? Provide comments as needed to explain your ranking. Example: "Data is being migrated into a new system and would require duplicative processes to publish at this moment" "Data has extensive quality issues and would require cleaning and preparation" "Data is already in a suitable, machine readable format for publication"

No

Free text

Overall Priority

What is the overall priority for the agency in publishing this data?
High - This data is of high value/import to the agency and is an immediate priority for publication
Medium - This data could be valuable if published, but is not an immediate priority
Low - This data, due to either readiness or value considerations, is not a priority at this time
No priority - This data cannot be published or is already published as an open data set and is discoverable on data.oregon.gov

No

High, Medium, Low, or No Priority

APPENDIX B: COMMONLY ASKED INVENTORY QUESTIONS

What is a “dataset”?

A dataset is defined as “a named collection of related records, maintained on a storage device, that contains data organized, formatted or structured in a specific or prescribed way” (ORS 276A.350). For the purposes of the Enterprise Data Inventory, a dataset should be considered a collection of data elements that are logically grouped together to perform a business function.

The Enterprise Data Inventory is focused specifically upon enterprise data assets/datasets that are directly related to the mission of the agency who collects, creates, maintains, aggregates, or otherwise utilizes them.

How granular should the data inventory be?

The Enterprise Data Inventory should contain datasets at a level that is granular enough to provide detail as to its use and business function, but not so granular that individual tables or data elements are being listed. See below for examples of granularity.

Dataset Name	Dataset Description	Level of Granularity
Oregon Case Management	Lists all permitting and case management for the agency	Too low – the dataset name and description are not clear enough to understand the purpose of the dataset or what information it may contain
Building Permitting Dataset	Contains information from the process of granting building permits, including location, business type, and business entity information	Appropriate level – the dataset and description identify at a high level what the data contains
Building Permitting data - Addresses	A database table of building addresses for the building permitting process	Too high – this may be an appropriately curated dataset for open data, but for the inventory this is too highly detailed
Building Permitting data - Business Entities	A database table of business entities provided by customers/clients as part of the permitting process	Too high – this may be an appropriately curated dataset for open data, but for the inventory this is too highly detailed

The “Building Permitting Dataset” encapsulates the addresses and entities elements within its description, so there is not an individual record or row needed for each table or possible dataset within the building permitting database or system. As

part of the agency’s Open Data Plan, specific datasets within the system may be listed or curated from the inventoried data for future publication, but they do not need to be identified as individual datasets within the Enterprise Data Inventory.

Should an agency inventory data they receive from, or report to, a federal program? What about data the agency purchases from a vendor?

If the agency creates, collects, maintains, or is mandated to care or control for the data, it should be included on the Enterprise Data Inventory. Data that the agency uses from other sources or receives from another public body who maintains custodianship of the record is not required to be included in the inventory. Examples of data that would be included and not included are below.

Data Type	Scenario	Include?
American Community Survey	The agency utilizes data available online from the American Community Survey to add median income to a map layer.	No. The agency does not contribute, collect, or create this data, so they do not need to include it in the Enterprise Data Inventory.
Workforce Innovation and Opportunity Reporting Data	The agency provides data to the federal government as part of an ongoing program.	Yes. The agency is responsible for collecting, creating, or maintaining this data as part of agency business, so it should be included in the Enterprise Data Inventory.
Sensor data purchased from a data broker	The agency pays another entity to collect this data and receives it as part of a contract.	Yes. The agency is paying for the creation or collection of this information and manages it as part of the agency’s mission, so it should be included in the Enterprise Data Inventory.

If an agency has level 4 data or classified data, should it be listed on the enterprise data inventory?

Data that may not be available for publication, sharing, or consumption to external parties should still be included on the Enterprise Data Inventory. The intent is to capture all enterprise datasets used by agencies in the support of their program areas, business units, and agency mission, regardless of whether it can be made publicly available or not.