Data Identification & Prioritization

Table of Contents

Section 1. Purpose

Section 2. State Data & Data Assets Defined

Section 3. Reasons for Publishing Data

Section 4. Identifying Data

Section 5. Assessing Impact

Section 6. Assessing Difficulty

Section 7. Rating Data

Section 8. Prioritizing

Section 1. Purpose

This document helps state agencies identify and prioritize data in order to produce a reasonable and workable plan for creating a comprehensive catalog of their data assets.

Section 2. State Data & Data Assets Defined

State data are items of information that are collected, maintained, and utilized by state departments and agencies for the purpose of carrying out State of Iowa business. State data are essential data required to conduct operations, and would include any data elements that are created, received, maintained, or transmitted.

State data assets are data that have been transformed into information that contains operational characteristics or ambient conditions for the purpose of communicating results, monitoring issues or problems, controlling processes and managing performance, improving operational effectiveness and efficiency, and/or facilitating actionable insights. Data assets would include datasets, as well as charts, maps, measures, stories and dashboard derived from such datasets.

Section 3. Reasons for Publishing Data

Publishing state data assets on the data portal facilitates:

- Improving the public's understanding of the cost and purpose of government services
- Improving governmental accountability and public participation
- Leveraging data held by different agencies by connecting datasets and finding new insights
- Eliminating redundancies by allowing the access of data in one place
- Improving decision making by better informing people with data
- Creating more efficient and proactive process for open records requests
- Encouraging innovative ideas (e.g., web applications) that enhance the lives of our citizens
- Increasing economic activity by generating new and rich content through new applications and services
- Complying with the requirements of the Taxpayer Transparency Act and Accountable Government Act.

3.1 Taxpayer Transparency Act

The Taxpayer Transparency Act, specifically Iowa Code Section 8G.3, makes publishing state data to the data portal all inclusive by defining agencies as a state department, office, board, commission, bureau, division, institution, or public institution of higher education – including elective offices of the executive branch, agencies of the general assembly, and the judicial branch. Programs and activities that are administered by or involve more than one agency are also included. The legislation requires agencies to provide information on performance outcomes related to funding actions or expenditures. Funding action or expenditure, according to the act, includes details on spending that is provided including but not limited to grants, contracts, and appropriations.

3.2 Accountable Government Act

The Accountable Government Act (AGA) (Iowa Code Chapter 8E) requires state agencies to disseminate performance plans, performance measures, performance targets based on performance data, and the performance data and data sources used to evaluate agency performance. It also authorizes the Department of Management access to agency records (i.e. the underlying data) relating to performance. Publishing data, other than that confidential under law, furthers this dissemination and offers new and more comprehensive ways to look at agency performance.

Section 4. Identifying Data

Identifying data state agencies collect, maintain or hold – even if it is historical data – is an excellent first step towards identifying datasets to publish on the data portal. Below are recommended activities to get started.

4.1 Agency Website

Reviewing data state agencies have already posted on their agency website is a logical first step. Look at Microsoft Excel files or public facing applications, which allow visitors to search for records. The data may not necessarily be accessible in bulk, or available through machine-readable mechanisms, but can serve as a good starting point.

Tip: Google search bar can facilitate finding data files on your website by entering the URL for your website and file type (e.g. site:www.educateiowa.gov filetype:xls).

4.2 Published Reports

Published reports are often populated with data which is compiled or aggregated from internal data systems. For example, a public report may indicate that an agency has closed 100 projects in the last month. The internal data system, which maintains information for each project, will likely have additional details that can be made public, such as, the type of project, its location, etc.

4.3 Public Information Officer

Agency public information officers will often have information regarding data or reports frequently requested by citizens or other agencies.

4.4 Statistical Analyses Performed

These analyses often use data from various sources, and are typically associated with issues or problems agencies deem important. The state agency's constituency may also find this information of value.

4.5 Federal Agency or Legislative Reports

Reports submitted to federal agencies or the state legislature (and their underlying data sources) can help identify data which can also be provided to the public. In addition, meeting

these reporting requirements (particularly statutory ones) might be accomplished simply by making the dataset(s) available on data.iowa.gov.

4.6 Exploring Data Published by Others

Look at data already published by other state agencies on the data portal. Explore the Open Data Network to see what others are publishing in different states, counties and cities, and the U.S. Government's data portal - data.gov. Both may provide a source for ideas.

Section 5. Assessing Impact

The more valued a dataset is, the higher the impact it is likely to have once published. The following offer considerations to use when assessing impact.

5.1 Strategic Initiatives

Data related to the Governor or state agency's strategic initiatives is of value if it can help demonstrate progress being made and/or if your agency's efforts are having the desired effect.

5.2 Agency Story

If the data helps improve the public's understanding of a state agency's mission and operations and/or quantifies results achieved by the state agency, it is of high value.

5.3 Insights on Key Issues

Data which helps explain issues or answer questions can be of great value to state agencies and their constituents.

High impact means data:

- Tracks strategic initiatives
- Tells agency story
- Is frequently requested
- Has a public impact
- Is in strong demand
- Is of timely interest
- Costs high $$$ to collect
- Provides an economic opportunity
- Facilitates reporting
- Encourages cross-agency collaboration

**Data does not have to meet all conditions to be considered high impact.

5.4 Frequently Requested

As demand is known and quantifiable, this should raise the value of data for publication. If the dataset is requested on a recurring basis, then state agencies may reduce duplication and obtain efficiencies by publishing data on the data portal.

5.5 Impact on Public

The data is likely of higher value if it is already apparent there is a deep impact and interest by the public.

5.6 Strong Demand from Key Interest Groups

The data might be of higher value to specific, narrow interest groups which may be a state agency's core constituency for those issues. It is important to not overlook these constituencies even if demand from the public overall is low.

5.7 Timely Interest

Announcements of progress or success or reactions to public criticism can be strongly supported by publishing related data, should it exist.

5.8 Cost to Collect and Maintain

If a state agency spends a great deal of money on a particular set of data, then it is highly likely that others would like to access it.

5.9 Economic Opportunity

In many cases, this will be unknown to state agencies in advance. Some of the greatest successes with making public data available have involved government data being commercially appropriated in useful ways. Any anticipated commercial use of state agency data should be taken into account.

5.10 Satisfy Required Reporting

Some required reports do not require extensive narration, and may be satisfied by publishing datasets alone, or in combination with interactive reports/summaries.

5.11 Facilitate Collaboration

Certain state functions may involve multiple agencies requiring access to similar data. State agencies may collect data that is of considerable value to another agency.

Section 6. Assessing Difficulty

Difficulty publishing data is often tied to how easily the data can be extracted and the quality of the data itself. The following offer considerations to use when assessing difficulty.

6.1 Structured Data

If data is contained in a fixed field within a record or file (e.g. contained in a database or spreadsheet), it will be much easier to publish compared to data in a document or paper format. Structured data typically has a data model that defines the fields data will be stored in and specifies how the data will be stored (e.g. data type – numeric, date, text, etc.). Structured data is much easier to extract from its data source(s) – thus making it easier to publish.

6.2 Missing or Incomplete Data

Missing or incomplete data prevents users from being able to effectively aggregate and compare values. If a dataset is missing relevant data values, it may be necessary to complete records from other sources (e.g. paper records or electronic documents). The more extensive or widespread the gaps in your data are, the more difficult it may be to publish your dataset.

Low Difficulty means data is:

- Structured
- Complete
- Unambiguous
- Consistent
- Non-redundant
- Based on standard references/protocols
- Free of confidential/sensitive data

**Data does not have to meet all conditions to be considered low difficulty.

6.3 Data Ambiguity

Data ambiguity arises when what the data represents is not precisely defined. This can lead to data values being misinterpreted. For example: DHS can represent two different government agencies, Department of Homeland Security at the federal level or Department of Human Services at the state level. Having to correct ambiguity in your data will make it more challenging to publish.

6.4 Data Inconsistency

Data inconsistency occurs where data values refer to the same thing but are recorded differently. For example, Mt. Pleasant and Mount Pleasant refer to the same town in Iowa, and Dept. of Public Health, DPH and Iowa Department of Public Health all refer to the same state

agency. However, since they are not recorded in a consistent way, values cannot be properly aggregated and compared. Correcting data inconsistencies can make publishing your data more challenging.

6.5 Data Redundancy

Redundant data usually occurs where data values for the same thing are recorded in multiple places. This could potentially lead to contradictions in the data. For example, if vendor addresses are entered on individual financial records, rather than in a unique vendor record, there is the potential for different addresses being recorded. When this happens, data users would not know which address is the correct one. Having to determine which data is correct will make publishing your data far more challenging.

6.6 Standard Reference for Measurement

Measured data lacking a standard reference method or measurement protocol lends itself to uncertainty, as it cannot be easily replicated. Additionally, if measured data lacked a standard and was collected by multiple individuals – the accuracy of the data becomes questionable. This is perhaps the most difficult data quality issue to deal with.

6.7 Confidential or Sensitive Data

If your dataset contains confidential or sensitive data (e.g. data protected by state law, such as Iowa Code Section 22.7 or other applicable Iowa Code section, or federal law, such as the Health Insurance Portability and Accountability Act, Social Security Number Protection Act, and Family Educational Rights and Privacy Act), it will be more difficult to prepare for publication. De-identification and other disclosure requirements can greatly increase the burden of publishing the data for public use if protocols and procedures have not been developed. Confidential or sensitive data in some cases could prevent a dataset from being published as a public dataset altogether.

6.8 Existing Processes

State agencies may be able to leverage existing processes to publish the data, such as exports for periodic department reviews, or routine exchanges of data with other agencies (e.g. data sharing agreements). It would also include any quality assurance processes to verify the quality and integrity of your datasets. Having such existing procedures in place may make the data easier to publish.

Section 7. Rating Data

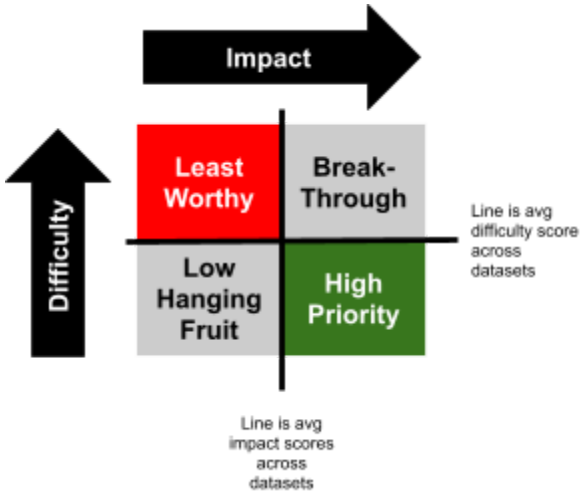
Rating can be done by an individual or as part of a team. For rating to be most effective it is important to have individuals involved who have a good understanding of the data the agency collects. Raters should give each dataset being reviewed a rating for impact and difficulty.

For impact, raters should give each dataset a score between 1 and 10, with 1 representing low impact and 10 representing high impact. For difficulty, raters should give each dataset a score between 1 and 10, with 1 representing data that is very easy to publish and 10 representing data that is very difficult to publish.

If multiple raters are used, then the dataset's impact and difficulty scores are based on the average of the scores given by raters.

Section 8. Prioritizing

Impact and difficulty scores are compared to the average impact and difficulty score for all datasets rated within the state agency to determine priority. So, once the impact and difficulty ratings are determined for each dataset, the average impact score and average difficulty score for all datasets will need to be calculated.



If the individual dataset score is above or equal to the average, the dataset is considered "High" for that category. If the individual score is below the average, the dataset is considered "Low" for that category.

High Impact, Low Difficulty datasets should be given the highest priority, as shown in green. Low Impact, High Difficulty datasets should be the last considered, as shown in red.