Open Data Audit and Publication Planning

Background

The concept of Open Data is about making data held by public bodies available and easily accessible online for reuse or redistribution at no or marginal cost. It has been defined by the Open Knowledge Foundation as *"a piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike".* Data Management and Open Data are key elements of a wide range of Public Service Reform Initiatives (see Figure 1).

The Minister for Public Expenditure and Reform launched the Open Data Initiative in July 2014.

Figure 1 – Linking Open Data to Public Service Reform Initiatives

Auditing to publication of datasets in the context of the Open Data Initiative

Ireland's Open Data Portal

One of the main priorities under the Open Data Initiative is the ongoing development, expansion and enhancement of the Open Data portal, https://data.gov.ie. An Open Data portal should contain official data that is available for re-use by all. Significantly greater availability of Open Data can play an important role in strengthening openness, transparency and accountability.

The Open Data portal links to datasets published by Government Departments and other public bodies. There is significant scope to build on the number of datasets on data.gov.ie by linking (through harvesting and where formats permit) to datasets already published on other portals and through the carrying out of data audits by public bodies with a view to publication in open format where appropriate.

Data Audits

It is important, to avoid such portals becoming 'data dumps', that a planned and structured approach is taken to the publication of data as open data taking into account the value, potential for re-use and contribution a dataset can make to achieve the economic, social and democratic benefits of Open Data. Organisational data audits play a key role in this regard. More generally, auditing of datasets[1] should be seen as part of an organisation's information management strategy. It is recognised that there is considerable value in having a comprehensive list of the datasets created, managed and maintained within Government Departments, regardless of whether or not they can be published at present.

To support the audit process, a Data Audit Tool has been developed at https://audit.data.gov.ie which can be used within each area to comprehensively audit and identify appropriate datasets. The Data Audit Tool is particularly useful where it is unclear which datasets are suitable for publication as open data. A list of the fields available on the Data Audit Tool is available in Appendix 1.

You should contact the Open Data Unit for access to the Data Audit Tool (opendata@per.gov.ie). There are two levels of access within an organisation. 'Manager' (who has overall authority within the organisation to approve datasets for publication) and 'User' (who can add datasets to the data audit tool but cannot approve them for publication). The Manager can add users to the data audit tool for their organisation (without having to contact the Open Data Unit) and view all datasets within their organisation which have been added to the data audit tool. A guide to using the Data Audit Tool is available on the portal. Approved datasets can be published directly on data.gov.ie via the Data Audit Tool.

In addition, the sample table below might be of use to support a basic data audit within an organisation.

Example of a Table showing summary of responses to a data audit.

No	Section	Dataset	Updated	Description of Dataset	Published?	Note
1	Name of Section/Division	Name of Dataset	Frequency	Summary description	Is this already published by the public body?	If published, provide location. In what format? E.g., PDF, Excel, csv. If not published online, provide reason.

Aim of Audit

The output of the audit would not only facilitate publication of datasets on the national portal https://data.gov.ie and the resultant potential benefits (economic, social and political) which could ensue, but would also promote the effective management and use / sharing of information across the Department. It would enable identification of:

- The extent and range of datasets that exist and are maintained (if in use)
- The ranking of datasets in terms of their importance to the delivery of Departmental objectives and the perceived gaps in useful data that might help the delivery of these objectives
- The potential for sharing datasets within the Department and the wider public sector
- The potential for publication and making available for re-use – obligations in this regard are set out in circular 12/2016 – 'Licence for re-use of public sector information'. It would be expected that greater publication and access to open data would reduce reliance on access to information under FOI. A guide for public bodies on the re-use of public sector regulations has been published by the Department of Public Expenditure & Reform.

---

[1] "Datasets" in the context of Open Data include databases, registers, ongoing data collections and surveys, geospatial datasets, and so on. Pure text-based information (emails, memos, reports, press releases) are not considered as "data" for this purpose.

Conduct of the audit/process

The steps involved in the audit / publication process include:

Audit
- Identification of datasets - the audit tool should be used for gathering the information on the audit. Personal data will not be published as Open Data unless irreversibly anonymised.
- The Open Data Audit Tool is an internal tool. Datasets that are classified as suitable for publication as Open Data are marked as 'approved' and can be automatically published to the Open Data Portal https://data.gov.ie
- Decisions on Publication – Senior Management in conjunction with their staff need to decide whether the datasets held, if any, are suitable for publication; or if datasets could be published in open format in the case where datasets are published already. In deciding whether to publish or not, the following should be taken into account - the likely demand, the current format/condition of the data, the frequency of updates, benefits and any risks

Publication Planning

In the context of the datasets included in the audit returns, it is proposed that the following steps be taken towards publication as Open Data. Each of the steps outlined below should take account of the balance between costs and anticipated benefits. Where datasets are hosted by one Unit using data from another, the publication planning process should involve both. Sample templates for a publication plan are included at Appendix 2.

In the context of developing a publication plan, the following needs to be considered:

1. Where a dataset is already published and linked to the Open Data portal, are enhancements or modifications required to make the datasets more searchable?

2. Where a dataset is already published, consider whether it can be linked to the Open Data portal. Are any enhancements required, e.g., in terms of formats, etc.

3. Whether unpublished datasets which are considered suitable for publication, can be published following any necessary actions (e.g., associating with Open Data licence, anonymisation, etc.), and publish as Open Data;

4. Where a dataset requires anonymisation before publication, discuss commencing this process. Can personal or commercially sensitive data be sufficiently anonymised or aggregated to allow publication?

5. Consider steps required to ensure compliance with the Open Data Technical Framework as outlined below:

   - Publish datasets in 3 Star format, at a minimum. The Open Data Technical Framework recognises the importance of publication in other formats. Public bodies may consider publication of datasets in multiple formats, for example, 1 Star (e.g. PDF), 2 Star (e.g. Microsoft Excel) in addition to the required 3 Star (e.g. CSV).
   - Licence - The Creative Commons Attribution licence, as recommended in the Technical Framework for Open Data, should be used.
   - Metadata - Use the recommended open data metadata format (DCAT) to describe each dataset
   - Data standards – When publishing Open Data, international standards defined by reputable standards organisations, such as ISO, the European Commission, W3C, IETF, OGC and OASIS should be used if possible. If international standards are unavailable or unsuitable, use

national standards. For specific topics such as geospatial, statistics, or health, use national standards as defined by the responsible organisation (OSI, CSO, HIQA, etc.).

- Unique Resource Identifiers - The Technical Framework sets out steps that will allow public bodies to achieve a minimum 3 Star Open Data. In the longer term, however, the intention is to progress to greater levels of linked data (4 and 5 Star). The use of Universal Resource Identifiers (URIs) is an important element of this longer term approach to Open Data and the Public Bodies Working Group will be tasked with developing a URI Strategy and agreeing a URI pattern for use under the Open Data Initiative, using international experience and best practice.

Resources

- Determine effort required to conduct audits and decide on which datasets are publishable.
- Some departmental areas may have very few datasets so audit and publication planning should not be particularly time-consuming. Other areas that hold large numbers of datasets, or where a lot of work is required to convert to the required formats, will be able to set out a plan to carry out this work over time.

Conclusion

The time/work commitment involved may not be particularly significant balanced against the benefit of carrying out a data audit (i.e. the Department would have a comprehensive audit of data holdings as a result promoting more effective use and data management as well as maximising the potential for publication).

Appendix 1: Data Audit Tool – Add a Dataset- Fields

- **Title**
- **Description**
- **Is there a demand for this dataset** (evidenced by FOI requests etc)? Yes/No
- **Who are the users of this dataset?** (eg academia, business, citizens, Departments and other public bodies etc)
- **What is the theme of the dataset?** (Agriculture, Arts, Crime, Economy, Education, Energy, Environment, Government, Health, Housing, Society, Science, Towns, Transport)
- **Please list keywords that describe this dataset?** (as comma separated values)
- **What geographical area does your dataset cover?** (eg international, national, regional, local)
- **What time period does the dataset cover?** eg 2014-2018
- **Is the dataset uses information from another source is it modified or aggregated by your public body prior to publication?** Yes/No
- **How frequently is the dataset updated?** (Never/daily/weekly/monthly/quarterly/biannually/annually/biennially)
- **When was the dataset created?** (in yyyy-mm-dd format eg 2012-10-28)
- **When did the last update take place?** (The last time the dataset was changed in yyyy-mm-dd format eg 2012-10-28)
- **Is metadata or descriptive information associated with this dataset?** Yes/no
- **Please indicate whether a standard metadata schema or an in-house schema is used.** (eg DCAT, ISO 19115, Dublin Core, Custom etc)
- **What is the Responsible Business Unit for this dataset?** (eg Irish Sports Council)
- **If the dataset is already published online please provide an URL to the webpage where the dataset is available.** (eg http://www.revenue.ie/en/about/statistics/receipts-taxhead.html)
- **For each data format that the dataset is available in (eg csv, xls, pdf, JSON, etc) please provide the direct download URL.** (option to add more than one)
- **If the dataset is not already published online, can it be published?** Yes/no
- **If no, please explain why not.** (eg it contains personal data or is excluded under copyright)
- **Is the dataset currently associated with a licence?** Yes/No
- **If yes, which licence?** (Creative Commons Attribution, Public Sector Information, Other Attribution, Other)
- **Could the dataset be published under the Creative Commons CC-BY 4.0 licence?** Yes/No
- **Would publication of the datasets held in your area deliver potential benefits to citizens, businesses, or the Public Service?** (Yes/No)
- **Are there gaps in the data?** (For example, a change in the system means that data was not collected in certain years, that data was only collected in certain years etc) Yes/No
- **Would publication require agreement from external bodies?** (EU, Copyright holders, other PBs etc) Yes/No
- **Is the data reliable – can we stand over a dataset's accuracy?** Yes/No?
- **Is anonymisation or aggregation of the datasets required?** Yes/No
- **If there are any other considerations please advise.**

CREATE

Appendix 2: Sample templates that could be used in an Open Data Publication Plan

Prioritisation

High Priority – initial linkage to data.gov.ie & work on enhancements
Medium Priority – link subject to demand
Not planned/Not publishable

Already Online – enhancements to improve linkage with data.gov.ie

No	Section	Dataset	Description of Dataset	Note	Proposed Action
					•
					•

Data Online – link to data.gov.ie and enhance openness

No	Section	Dataset	Description of Dataset	Note	Proposed Action
					•

Not Online, but can be published

No	Section	Dataset	Description of Dataset	Note	Proposed Action
					•

Not Online – Publication not currently planned

	Section	Dataset	Description of Dataset	Note