Data and Digitalisation Steering Group

Data Triage Playbook

May 2021 | Version 1

Table of Contents

Introduction

About ENA

Energy Networks Association (ENA) represents the owners and operators of licenses for the transmission and/or distribution of energy in the UK and Ireland. Our members control and maintain the critical national infrastructure that delivers these vital services into customers homes and businesses.

ENA's overriding goals are to promote UK and Ireland energy networks ensuring our networks are the safest, most reliable, most efficient and sustainable in the world. We influence decision-makers on issues that are important to our members. These include:

- Regulation and the wider representation in UK, Ireland and the rest of Europe
- Cost-efficient engineering services and related businesses for the benefit of members
- Safety, health and environment across the gas and electricity industries
- The development and deployment of smart technology
- Innovation strategy, reporting and collaboration in GB

As the voice of the energy networks sector, ENA acts as a strategic focus and channel of communication for the industry. We promote interests and good standing of the industry and provide a forum of discussion among company members.

Our members and associates

Membership of Energy Networks Association is open to all owners and operators of energy networks in the UK.

- Companies which operate smaller networks or are licence holders in the islands around the UK and Ireland can be associates of ENA too. This gives them access to the expertise and knowledge available through ENA.
- Companies and organisations with an interest in the UK transmission and distribution market are now able to directly benefit from the work of ENA through associate status.

ENA members

Cadent – Your Gas Network
electricity north west – Bringing energy to your door
ESB NETWORKS
gtc
nationalgrid
nationalgridESO
Northern Gas Networks
Northern Ireland Electricity Networks
NORTHERN POWERGRID
Scottish & Southern Electricity Networks
SGN – Your gas. Our network.
SP ENERGY NETWORKS
UK Power Networks – Delivering your electricity
WALES&WEST UTILITIES
WESTERN POWER DISTRIBUTION – Serving the Midlands, South West and Wales

ENA associates

- Chubu
- EEA
- Guernsey Electricity Ltd
- Heathrow Airport
- Jersey Electricity
- Manx Electricity Authority
- Network Rail
- TEPCO

Introduction to Data Triage

The Energy Data Taskforce recommended that organisations across the energy sector should embrace the principle of "Presumed Open" which encourages data owners and custodians to make datasets as open as possible. This is intended to unleash innovation, improve efficiency, and encourage new business and technological solutions which are essential for the UK to meet the Net Zero target.

Presumed Open is the principle that data should be as open as possible. Where the raw data cannot be entirely open, the data custodian should provide objective justification for this. For data to be made 'as open as possible', it is necessary to have a formal process which can be used to identify potential issues and mitigate them as necessary, this is referred to as Open Data Triage.

Open Data Triage is a process to systematically identify issues (Privacy, Security, Commercial, Negative Consumer Impact or Legislation and Regulator Barriers) with a dataset which limit their potential openness and then identify what techniques can be used to mitigate these issues.

This document is designed to act as a framework to support organisations' Data Triage process to provide a consistent approach throughout the energy utilities industry, where there is significant commonality in datasets but not currently the approach to making these openly available.

This Playbook is designed to act as a facilitator from the point on where the following two conditions occur:

- An open data request has been instigated by a third party through a formal or information mechanism;
- **T**he need to identify datasets to support presumed open internal to an organisation has been triggered either through an ad hoc request or a proactive data openness approach.

Figure 1 provides a high level flow of the Data Triage Playbook process described throughout this document.

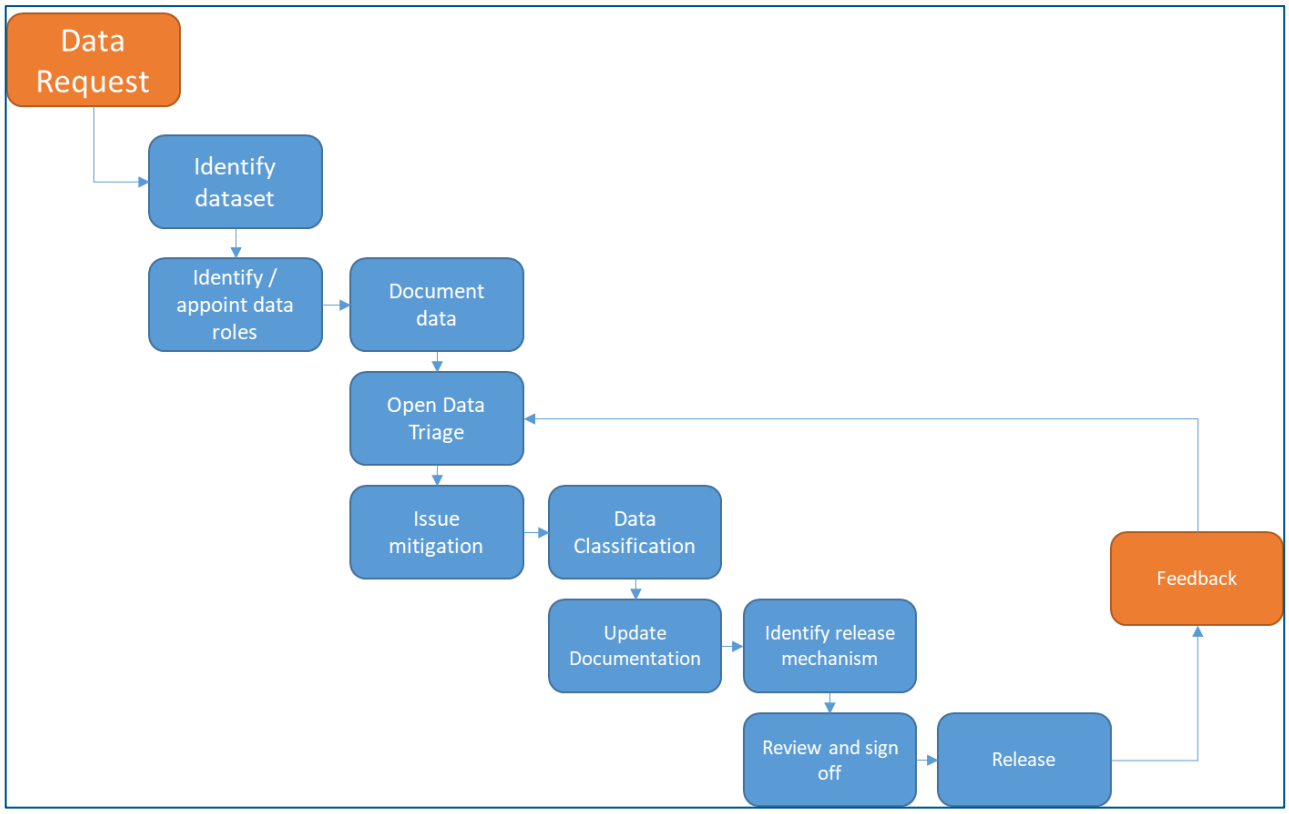

Figure 1: High level Data Triage Playbook process

This playbook is designed to support a consistent approach to the data triage process across ENA members to provide commonality in the assessment of data across the four defined data classifications:

- Open: Data is made available for all to use, modify and distribute with no restrictions;
- Public: Data is made publicly available but with some restrictions on usage;
- Shared: Data is made available to a limited group of participants possibly with some restrictions on usage;
- Closed: Data is only available within a single organisation.

Identify Dataset

The first stage of the assessment is to identify if / where a usable dataset or how to generate such a dataset that appears to provide the basis to meet the requests' need. A 'usable dataset' is a discrete collection of data which relates to a focused, coherent element but provides enough information to be of practical use.

A usable dataset is likely to be centered around:

- Data source (device, person, system);
- Subject of data (technical, operational, personal, commercial);
- Time and granularity (collection period, frequency of data collection, inherent aggregation);
- Location (country, region, public/private area);
- Other logical categorisations (project, organisational department, etc.).

This is a critical step as it would be infeasible to try and identify and mitigate all potential issues relating to a large database or system in a reliable way. Breaking the problem down into small pieces and considering each dataset independently makes this much more achievable.

At this stage thought should also be given to the ease of data extract and processing. Some systems may have preconfigured data exports or reports which are generated for internal use, however, many of these are likely to be focused on meeting a specific business need.

Where a dataset is not readily available, i.e. from an existing compiled dataset or easily extractable from a relational database or other system thought should be given to the time required to facilitate the extract of the dataset. Consideration should be given to this in relation to the priority of the data (internal prioritisation or other) and a set of rules should be identified based on priority levels. An example of this is provided below.

Table 1: Dataset priorities by time to extract

Dataset Priority	Maximum time to extract (Hours)
Low	8
Medium	16
High	32

Where it is determined that the dataset cannot be extracted in the time allowed based on its priority this should be robustly captured, documented and shared with the requester (as a minimum) following approval by the Data Owner and Data Manager as required.

Identify / Appoint Internal Data Roles

For each dataset identified it is key to identify or where not previously identified, appoint internal data roles to support the next phases and on-going data needs. The following section provides an overview of the skills and / or capabilities required for each stage of the process.

Table 2: Process Stage Responsibilities

Process Stage	Responsibilities
Document Data	• An individual with the understanding of the data, its origins and lineage to suitably describe the data • A person or team who understands how metadata and data dictionaries should be used to suitably describe the data
Open Data Triage	• An individual who understands the data's contents and meaning to make an informed decision on the triage questions • A person or team who manages the triage process for the organisation
Issue Mitigation	• Understanding of the specific issue that needs mitigation and the reasoning as to why • A person or team with the technical skills and capability to facilitate the mitigation requirements of the data
Data Classification	• An individual who has the ownership / responsibility for the data • A data focused person or team who manage the data triage process to ensure a level of consistency with the data classification process
Identify Release Mechanism	• A person or team responsible for the release of data to ensure a consistent approach based on the requester, triage and classification results
Review and Sign Off	• A person or persons with the required authority to enable data to be released internally or externally

To support an outline of example role descriptions are included here:

Owner

A Data Owner is identified as the specifier of the business requirements of the data and its quality. The data owner role is usually assigned to a Senior Manager in the respective business function or department. They shall also be accountable for the data itself.

Curator / Operator

Data Curators / Operators typically operate the data life-cycle based on the defined standards. They create and maintain this data. This role is commonly taken by staff in the respective department of the relevant Data Owner, or in dedicated support functions.

Steward

A Data Steward supports the business departments in the desired use of data. Their role is routinely organised by data domains (e.g. customer data, asset data and system data). Data stewards evaluate requirements and problems with data, and support projects and digitalisation initiatives as experts for their respective domain.

Data Scientist

A Data Scientist typically supports the data mitigation and obfuscation phase, utilising data science techniques to facilitate the release of the data without compromising the security of private or personal information.

Data Process Administrator

A Data Process Administrator is responsible for managing the overall Data Triage process, collating and storing the relevant information for review and audit purposes.

Data Manager / Data Governance Group

A Data Manager owns the Data Triage process and is likely to be the approval of all or the majority of datasets for release.

Documenting the data

Key to understanding the data being triaged, and ultimately shared, is the data describing it. As a minimum the metadata, as described below, should be captured and it is good practice to capture data dictionaries.

Metadata

Once a dataset, where it doesn't currently exist, and data roles are identified the data should be robustly documented. The first stage of documenting the data sets is to produce metadata. Metadata is a dataset that describes and gives information about another dataset. This can be used by potential data users to understand what they may be able to do with the data technically and legally.

In line with the Energy Data Best Practice Guidance, it is proposed to use metadata based on the Dublin Core standard. The metadata structure for this project is as follows:

Table 3: Dublin Core Metadata Structure

Element	Description	Dublin Standard
Title	Name given to the resource	Core
Creator	Entity primarily responsible for making the resource	Core
Subject	Topic of the resource *(e.g. Keywords from an agreed vocabulary)*	Core
Description	Account of the resource	Core
Publisher	Entity responsible for making the resource available	Core
Contributor	Entity responsible for making contributions to the resource	Core
Date	Point or period of time associated with an event in the lifecycle of the resource	Core
Type	Nature or genre of the resource such as a data group	Core
Format	File format, physical medium, or dimensions of the resource	Core
Identifier	Compact sequence of characters that establishes the identity of a resource, institution or person alone or in combination with other elements *e.g. Uniform Resource Identifier (URI) or Digital Object Identifier (DOI)*	Core

Source	Related resource from which the described resource is derived *(e.g. Source URI or DOI)*	Core
Language	Language of the resource *(Selected language(s) from an agreed vocabulary e.g. ISO 639-2 or ISO 639-3).*	Core
Relation	Related Resource *(e.g. related item URI or DOI)*	Core
Coverage	Spatial or temporal topic of the resource, spatial applicability of the resource, or jurisdiction under which the resource is relevant	Core
Rights	Information about rights held in and over the resource such as Open Licence	Core

Examples of completed Metadata

Error! Reference source not found. Table 4 and Table 5 demonstrate two implementations of all or part of the Dublin Core metadata standard.

Table 4: Transformer detail for the South West Licence Area	
Element	**Description**
Title	Transformer detail for the South West Licence Area
Creator	Western Power Distribution
Subject	Transformer; Rating; Power; Impedance
Description	Key technical information for the transformers in WPD's South West region, including but not limited to voltage, impedance and ratings
Publisher	Western Power Distribution
Contributor	Data & Digitalisation
Date	2020-12-03 08:38:00 (UTC)
Type	System and Network
Format	CSV
Identifier	TX_South_West

Source	Western Power Distribution
Language	EN
Relation	LTDS
Coverage	South West
Rights	WPD Open Data Licence

Table 5: Embedded Wind and Solar Forecasts	
Element	**Description**
Title	Embedded Wind and Solar Forecast
Creator	National Grid Electricity System Operator
Subject	Embedded; Forecast; Generation; Solar; Subscribable; Wind
Description	Electricity System Operaor (ESO) publishes at a half-hourly resolution the embedded wind and solar forecast from within day up to 14 days ahead. This forecast get updated to an hourly basis
Publisher	National Grid Electricity System Operator
Contributor	National Grid Electricity System Operator
Date	2019-11-07 09:24:00 (UTC)
Type	Generation
Format	CSV
Source	National Grid Electricity System Operator
Language	EN
Rights	National Grid ESO Open Data Licence

Data Dictionary

A data dictionary in this instance is a repository that contains records about the elements and fields within the data. For each field within the dataset it is recommended that the following should, as a minimum, where applicable be captured:

Table 6: Data Dictionary Structure	
Element	**Description**
Title	Name of field *(e.g. Rating, Name, ID)* [Power Rating]
Type	Text, Numeric, Date/Time, Binary, Geometry [Numeric]
Description	Account of the field [The *standard* rating of the asset as defined in IECXXX]
Comment	Ad hoc usual information to increase dataset understanding [data is correct as of 2019]
Example	An example of the data contents [35.5]
Unit	Capture of the International System of Units or other for the field [MW]

Examples of completed Data Dictionary

Table 7 and Table 8 demonstrate two examples of data dictionaries provided for fields within the dataset described in Table 4 and Table 5 respectively.

Table 7: Data Dictionary for Settlement Date field [data described in Table 4]	
Element	**Description**
Title	Settlement Date
Type	Date
Description	The period from 00:00 hours to 24:00 hours on each day
Example	24/03/2020

Table 8: Data Dictionary for Embedded Solar Capacity field [data described in Table 5]	
Element	**Description**
Title	Embedded Solar Capacity
Type	Integer
Description	Estimated Embedded Solar Capacity. This is National Grid ESO's best view of the installed embedded solar capacity in GB. This is based on publically available information compiled from a variety of sources and is not the definitive view. It is consistent with the generation estimate provided.
Example	13080
Unit	MW

Open Data Triage

The triage process must consider elements such as privacy, security, commercial and consumer impact issues. Where the decision is for the raw data to not be made open the Data Owner / Curator must share the rationale for this and consider sensitivity mitigation options (data modification or reduced openness) that actively endeavours to maximise the usefulness of the data. Where a mitigation option is implemented the process followed to provide the mitigation should be made publicly available with reference to the desensitised version of the data (likely in the metadata description or data dictionary comments if only relevant to a specific field). In the cases where no data can be made available then the rational should be robustly documented and made available for review and challenge as part of the feedback process.

Figure 2 shows a high-level representation of the proposed process outlined.

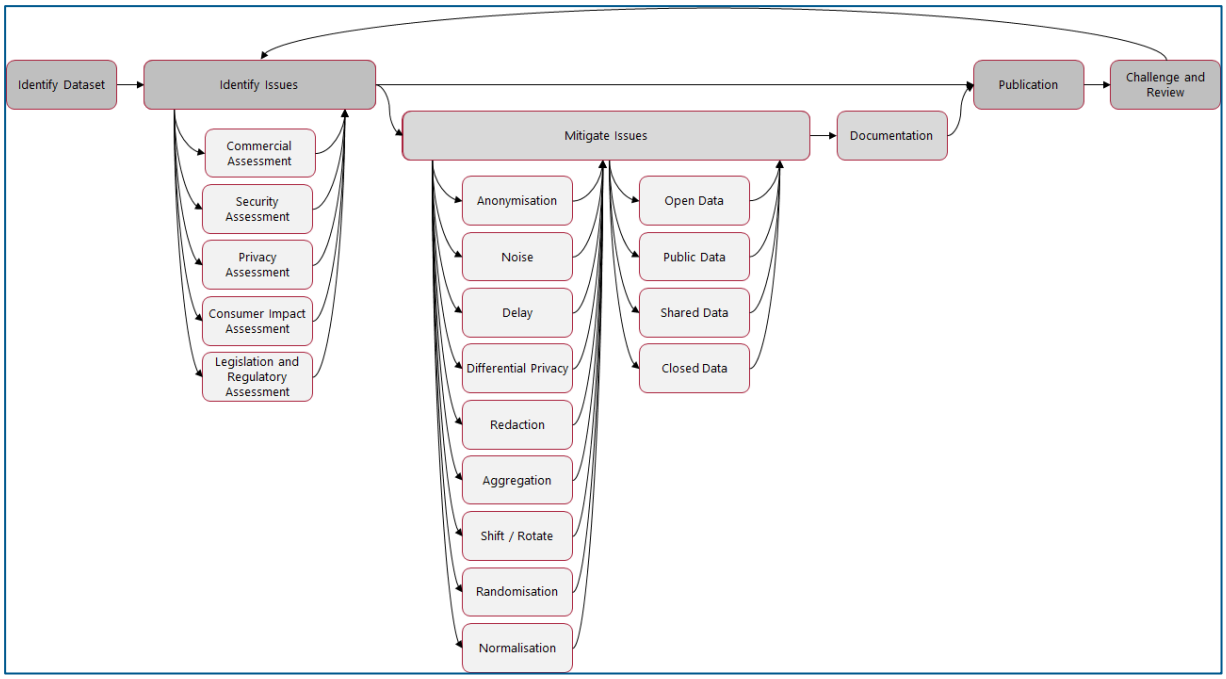

Figure 2: Detailed Open Data Triage process

Identifying issues

In order to identify issues with sharing the data in its current format and content a small number of key questions should be answered in respect of the dataset. It is proposed that the Data Owner will answer these questions with the support of the organisation's data governance / administration team. An example of questions, developed by Energy Systems Catapult, is provided below.

Table 9: Identifying Issues

Risk Factor	Considerations	Further Guidance
Rights	Does the organisation own or have a legal right to publish the data? Has the organisation (or a contracted third party) generated or collected the data?	In some cases where data has been shared with the organisation by the data owner, there may be rights (via data licencing) or obligations (e.g. Environmental Information Regulations or DCUSA Embedded Capacity Registers) which provide legal grounds for publication of the dataset or derived dataset. Where a dataset combines owned and shared data this should be recorded as shared data.
Privacy	Does the dataset contain personal data?	Consider if sharing this dataset infringes the General Data Protection Regulation (the GDPR) or the Data Protection Act 2018.
Security	Does this dataset contain information that creates an incremental security risk or exacerbate an existing risk?	Consider if the dataset contains information that is not already public by another source. e.g. existing organisation reports, third party datasets (e.g. satellite imagery), etc. If an alternative source exists then consider if the organisation's dataset contains more detailed data (granularity, latency, accuracy, etc.). Where the dataset contains novel information, it should be assessed for new security issues such as: security of supply (Distribution Code - Engineering Recommendation P2), security of assets, security of people, etc. Where the dataset replicates or extends existing public datasets, consideration should be given to the incremental risk publication would create. For example, does the increased granularity of data create more risk?
Commercial	Does the dataset contain information that could damage the commercial interests of the organisation?	Data Best Practice Guidance describes commercial information as "Data that relates to the private administration of a business or data which was not collected as part of an obligation / by a regulated monopoly and would not have been originated or captured without the activity of the organisation". An example of this is Section 105 of the Utilities Act 2000, which states that 'information

		relating to the affairs of any individual or business, obtained under or by virtue of the Act (or under Part I of the Gas Act 1986 or Part I of the Electricity Act 1989 Act or the Energy Acts) shall not be disclosed, save where permitted under the Act'. Commercial issues could also arise due to prior data licencing which would preclude sharing the same data under different terms (e.g. exclusivity agreement, confidentiality agreement), data which is (or could be interpreted as) inconsistent with regulatory reporting or could be seen as exploiting the organisation's licenced role (e.g. could undermine the Competition Act 1998, Enterprise Act 2002 or Enterprise and Regulatory Reform Act 2013).
Negative Consumer Impact	Would sharing this dataset result in a negative impact on consumers, concerned with commerciality and ethics?	Consider if access to this data is likely to drive actions, intentional or otherwise, which will negatively impact the consumer. For example, data about procurement assessment processes could drive up prices for products or services which increase costs for consumers.
Data Quality	Is a dataset quality caveat required for end users?	Data quality is subjective. A dataset may be perfectly acceptable for one use case but entirely inadequate for another. Data accuracy can be more objective but there remain many instances where the required precision differs across use cases. Data quality should not be seen as a reason for not sharing as potential users may find the quality acceptable for their use, find ways to handle the quality issues or develop ways to solve issues which can improve the quality of the underlying data. However, known limitations (quality or accuracy) should be clearly documented and where there are uncertainties a robust quality disclaimer should be used.
Other	Are there any other concerns associated with sharing this dataset not covered in the responses to the questions above?	If there are any other issues that require mitigation before publishing the dataset that haven't been addressed by the questions above. For example, specific legislative or regulatory barriers.

Summary

Providing a summary of the selected question set using a RAG status will inform on the required next steps, as shown in the example below.

Table 10: Identifying Issues – Example Questions Summary			
Risk Factor	**Initial Risk**	**Mitigation**	**Residual Risk**
Rights	GREEN		GREEN
Privacy	GREEN		GREEN
Security	GREEN		GREEN
Commercial	GREEN		GREEN
Negative Consumer Impact	GREEN		GREEN
Data Quality	GREEN		GREEN
Other	GREEN		GREEN

Mitigating issues

Where issues with sharing the data in its current format are identified mitigation techniques should be employed. Dependent on the specific issue different mitigation techniques should be employed. A number of potential open source implementation techniques for manipulating data and removing/concealing sensitive data are outlined here.

The aim of these techniques is to anonymise the data and protect what is private or personal in data sets whilst retaining as much of the important features/information for the possible use cases that the data is intended for. Removal or modification of too much data, or types of data, may make it unfit for the intended purposes. In contrast, adjusting the data too little may result in private data being insufficiently anonymised.

The techniques presented here typically fall into following categories:

- Modifying the original data to hide the private features;
- Removing features that are strongly linked to the private identifiers;
- Creating new features which retain the important elements of the data but remove the private features.

  Aggregating data for example is one way to do this.

A random element is often an important element in any of these techniques. If any procedure is deterministic then it is easier to reverse engineer and hence reveal the protected information. The process involves more than just removing unique identifiers which relate directly to an individual. In addition, there are also so-called quasi-identifiers which may be strongly correlated with the identity of an individual.

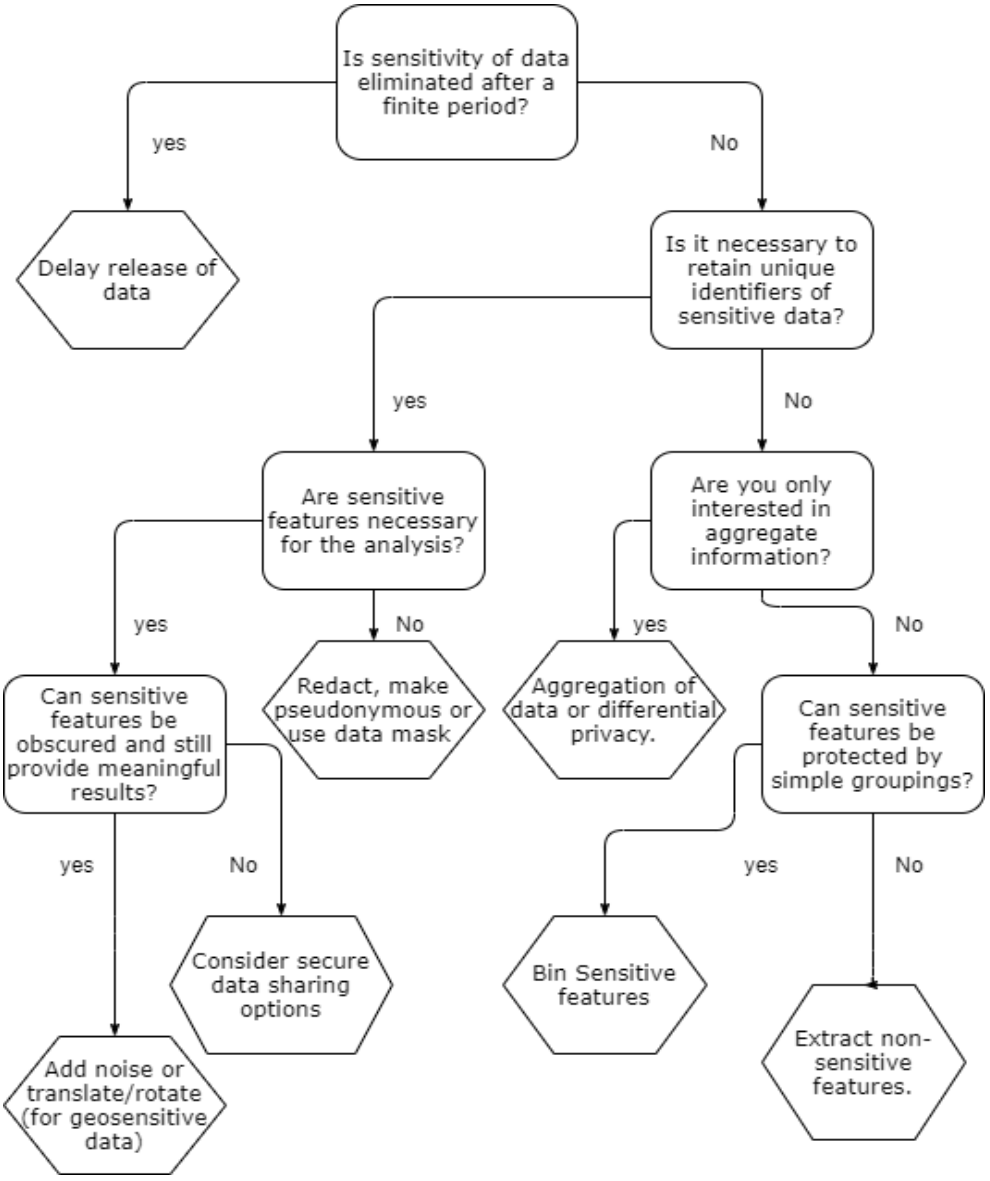

Figure 1: A simple flow diagram to show how sensitive data may be properly anonymised for different types of use cases

The use case (highlighted at data request phase) will support determining what a suitable level of manipulation is. For one use case the same modification will make a data set useless for the intended purpose but for another it could be entirely suitable. For example, aggregating smart meter data of consumers is a useful technique to conceal an individual households' demands. If the aim is to forecast the demand on the LV network this is a valid technique, and the aggregation does not hinder the process implemented. However, for determining energy efficiency interventions for individual consumers this is clearly inappropriate as it hides the useful behavioural information. In the latter case perhaps, derived features would be a sufficient solution.

Determining appropriateness of mitigation

Where a dataset requires a form of mitigation thought should be given to the time required to facilitate this mitigation to enable the dataset to be effectively shared. Consideration should be given to this in relation to the priority of the data, the current risk level associated with the dataset from initial Data Triage, and a set of rules should be identified based on priority levels. An example of this is provided below.

Table 11: Dataset priorities by time to extract

Dataset Priority	Risk Level	Maximum time to extract (Hours)
Low	Amber	4
	Red	8
Medium	Amber	8
	Red	16
High	Amber	16
	Red	32

Where it is determined that the dataset cannot be mitigated in the time allowed based on its priority and risk level this should be robustly captured, documented and shared with the requester (as a minimum) following approval by the Data Owner and Data Manager as required.

Overview

This section describes some of the major data manipulation categories, how they work and some open source implementations that are available.

Table 12 provides a summary of the techniques. The following subsections give some further information on each of these techniques.

The methods often overlap in the functions they perform. For example, redaction often means the data is deleted but can also mean that it is edited and is not identifiable. This is similar to pseudonymisation which replaces identifiable information with an artificial identifier.

Table 12: Summary of some modification techniques used to remove sensitive information from a dataset.

Technique	Description	Implementation
Redaction	Removing or overwriting selected features	https://pypi.org/project/piianalyzer/ for identifying PII in datasets. Python script for redacting PDFs: https://github.com/JoshData/pdf-redactor and also https://github.com/madisonmay/CommonRegex
Pseudonymisation	Replacing identifying features with a unique identifier that retains the reference to an individual whilst breaking the link with the `real world' identity	Python: https://github.com/fzaninotto/Faker And see: https://github.com/elastic/anonymize-it which is built on Faker. The digest package in R creates hashes for IDs which must have identifiers modified: https://cran.r-project.org/web/packages/digest/index.html.
Noise	Combining the original dataset with random data to conceal features of the data	Various packages can generate random noise. E.g., In Python both Gaussian random numbers and random categorical values can be generated by the NumPy package: https://numpy.org/. In R similar functions exist including: rnorm (Gaussian distribution) https://www.rdocumentation.org/packages/compositions/versions/1.40-5/topics/rnorm, or sample (positive integers): https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/sample
Delay	Deferring publication of data for a defined period.	N/A
Differential Privacy	An algorithm or model which obscure the original data to limit re-identification.	Google have developed an open source set of libraries for various differential privacy algorithms: https://github.com/google/differential-privacy

		IBM differential privacy tool (based on python): https://github.com/IBM/differential-privacy-library
Data Masking	Process for hiding original data with modified content.	PostgreSQL has a number of masking functions: https://postgresql-anonymizer.readthedocs.io/en/latest/masking_functions/ Faker in Python can also be used for data masking: https://github.com/fzaninotto/Faker
Aggregation	Combining data to reduce granularity of resolution, time, space or individuals	Python's pandas package, https://pandas.pydata.org/, contains a wide range of aggregation functions including mean, median, max, min, etc. There is similar functionality which comes as standard in R.
Translation/ Rotation	Altering the position or orientation of spatial or time series data	Some implementations can be achieved using random noise generation (See the Noise row above)
Feature Extractions or Engineering	Extract or generate new features from the data which hide the private data and replace	Dependent on the use case but can be implemented using similar techniques such as aggregation or binning.
Data Bucketing/Binning	Process the data into groups.	Most programming languages have methods for quantising data. In python for example there are the cut and qcut functions in the pandas package: https://pandas.pydata.org/.

Pseudonymisation

One of the simplest ways to remove personally identifiable information (PII) is to simply replace it with an artificial identifier e.g. a name can be replaced with an integer ID. In the case of multiple fields with PII (e.g., a name and address) then the combination of these fields can be replaced with a unique identifier. The remaining data retains the features in the data without revealing the individuals. Since unique identifiers are used, the original identity can then be easily found if this information is available. The ability to link back to the original data makes it useful but also carries risk if the data linking the identifiers to individuals is obtained. Hence sufficient measures are required to keep this data separate.

The main challenge for pseudonymisation now comes down to identifying what is PII. This is not always obvious since something that is not usually PII could be in different contexts. For example, depending on the size of the data, age may not be an identifying characteristic if there are sufficient entries to mask the individual. If not, then the data is susceptible to an inference attack, where an individual can be inferred by various data mining techniques.

There are many open source tools which can be used to perform pseudo-anonymisation. In python a popular package is the faker library[1]. This tool can produce artificial names, addresses, email addresses and text, amongst other things, which can then be used to replace the PII information in the document of interest.

Alternatively, creating "fake" versions of the data may not be desirable and instead the data can be replaced with random integers. Generating such random numbers is generally trivial in any open source programming language.

Redaction

Redaction refers to editing data. Redaction often refers to documents which require sensitive data to be removed although a table or database could also be redacted.

One of the simplest ways to remove sensitive data is to simply remove it from the dataset i.e., if one column of a data set has names of individuals then this field could be deleted. Alternatively, they could be replaced with other values. In the latter case this could be seen as similar to pseudonymisation as long as the replacement information uses a unique identifier to ensure that individuals are distinguishable.

Redaction can be applied to several different types of data, i.e., pdf's etc. In this case the personal identifiable information must be identified before it can be replaced. This can be impractical to do manually but there are packages which can be used to simplify the task. PII Analyzer searches [2] common regular expressions to identify certain types of PII such as email addresses, credit card numbers, phone numbers etc. There are also other open source tools such as pdf-redactor [3] which have been used to redact data from PDFs. Of course, a user could also write their own PII identifier by utilising regular expression datasets. Once the PII has been identified these can be replaced.

Noise

One way to scramble a dataset is to add noise. This can be added to all forms of data sets. For image data, noise can be added to each pixel to obscure the picture, or for a time series adding noise can reduce distinguishable features. In energy systems one use case would be to add noise to smart meter data to hide particular behaviours. There is a trade-off between the variability of the noise added and the reduced utility of the data set. High variability means any useful features from the data set may be completely hidden but so is any private information. In contrast low levels of noise may not hide any private information but the data may have higher utility for any algorithms deployed.

Since generating random data are core functionalities for most programming languages, it is relatively easy to generate and update a dataset with random noise. Choosing the right type and level of noise to add will depend on the type of data you are using, and the level of privacy required. For example, for time series data simple white noise may be suitable. The standard deviation (a measure of spread) of the data could then be chosen based on the frequency of the components that the user wishes to conceal.

Differential privacy, below, also involves a more sophisticated method for adding randomness to a dataset.

Delaying Publication

[1] https://github.com/fzaninotto/Faker
[2] https://pypi.org/project/piianalyzer/
[3] https://github.com/JoshData/pdf-redactor

When data has a limited time sensitivity, for example location data (to understand traffic behaviour say) being made immediately public would not be appropriate as it could be used to track someone in real time and perhaps identify the individual. However, releasing this data at a later point can not only reduce this risk but further data manipulations can be applied to make the data even more secure (such as removing potential home and work identifiers of an individual, see Section on Translation/Rotation).

Differential Privacy

Differential privacy allows collection of information without revealing private data. Even anonymised data can be used to identify individuals via linkage attacks were grouped non-private data can be used to identify individuals. Differential privacy removes the possibility of linkage attack and still allows aggregate insights and statistics without revealing individual's information. The goal is to ensure that the removal of an individual from the data set should not affect the statistical functions that run on the data. The inserted randomness in the data should be dependent on the size of the dataset (smaller the data, the more randomness needed to conceal identities). If there is information on how the randomness is added to the data then the statistical estimations can be made which are accurate representations of the original data[4].

Issues with differential privacy are relatively complex and hence open source libraries are often necessary rather developing your own implementations. Google uses differential privacy libraries to collect data for e.g., their maps tools to say how busy various businesses are during their day [5]and have made some of libraries available. Similarly, IBM have some differential privacy mechanism and models they have made available[6].

Data Masking

Data masking is a way of hiding confidential data such as credit card numbers, names etc. by using modified content whilst trying to ensure the data remains useful/meaningful. Some common techniques are:

0. Substitute values;
1. Encrypting the data with a key;
2. Shuffling the data within the same field/column;
3. Deleting data;
4. Scrambling the characters.

Aggregation

Aggregation refers to grouping of data. This could be the summing, or averaging data, or even clustering individuals. For the purposes of anonymising data, aggregation can be useful since general statistics can be generated which conceal the individuals or remove private features from the data. Aggregation can also refer to aggregating across spatial levels and/or temporal.

[4] https://accuracyandprivacy.substack.com/p/differential-privacy-an-easy-case
[5] https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html
[6] https://diffprivlib.readthedocs.io/en/latest/

To take smart meter data as an example again. In its raw half-hourly form, there may be many private features of a households day-to-day behaviour exposed. However, aggregating up to daily or weekly level, conceals all this behaviour. However, it may not conceal when a household is on holiday and the home is empty. In this case, aggregating across a street may be more advantageous for analysis. This both conceals the individual's behaviour and makes it unlikely to identify individuals who have gone on holiday (unless many of the occupants go on holiday at the same time).

Translation/Rotation

Transformations can be performed on geospatial data to remove sensitive geographical information. For example, GPS data may reveal where a person lives or works. Applying translations (i.e., horizontal shifts) or rotations to the dataset can help conceal the journey of an individual's GPS locations. As with most of the anonymisation techniques described, a random element must be introduced to the data. If the same translations and rotations are applied to all the datasets, then it can be relatively easy to reverse engineer the data and the locations.

Further anonymisation may be required in addition to these transformations. For example, it may be possible to identify the location by linking the journey path to road maps and thus reverse engineer the data despite the randomisation. In this case one solution would be to delete an area around the origin and destination to at least conceal these locations.

If the data is stored as longitude and latitude co-ordinates, then the translations can be simply applied to these values.

Feature Extraction/Engineering

Instead of sharing the full dataset it may be worth only sharing particular features of the data set. These could be the aggregated statistics as mentioned in the Aggregation section or other user defined values. As an example, consider EV usage. Instead of sharing the entire charging data of a user, the data could be reduced to some useful attributes such as average daily charge, maximum/minimum charge, and variance of daily charge.

To conceal the data even further, these features could be clustered into a finite number of groups and instead of sharing the reduced features, representatives from each cluster could be shared instead. See Data Binning below.

Data Bucketing/Binning

Certain information like ages and town of birth can be used as quasi-identifiers. By grouping data into particularly 'buckets' or 'bins' the risk of using this information for identifying individuals can be reduced. For example, ages can be put into specific age bands and town of birth could be grouped into regions or counties.

A more advanced way to group individuals is to use supervised discrimination methods, also known as clustering methods, such as k-means and Gaussian mixture models. These methods are commonly included in most scientific programming languages. For example, Python can implement both k-means and Gaussian mixture models as part of its scikit-learn package: https://scikit-learn.org/stable/

Reassessing issues

Once the data is obfuscated, delayed, or otherwise mitigated the initial issue identification question answers should be re-reviewed, as described in the 'Identifying issues' section and rescored following the RAG status. This in turn can then be used to define an appropriate data classification.

Data Classification

Data classification is required to provide a common understanding of data, its content sensitivity and availability to be shared externally.

Energy Systems Catapult's (ESC) Energy Data Taskforce Report[7] utilises the Open Data Institute's (ODI) Data Spectrum shown in Figure 2.

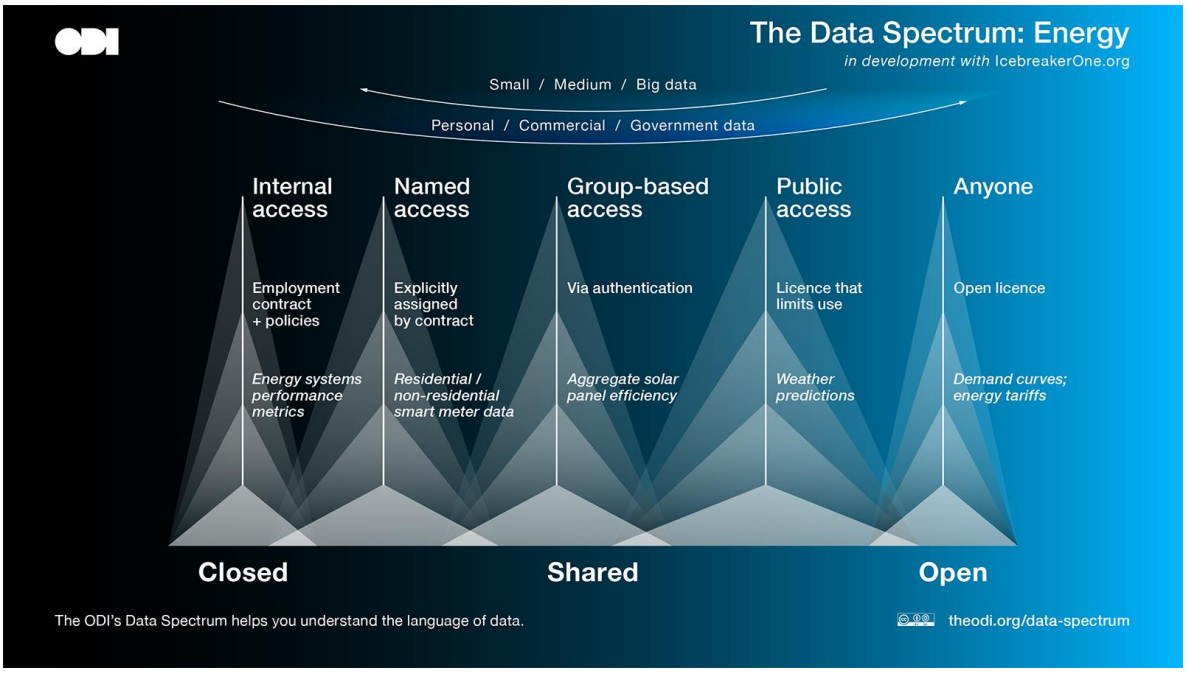

Figure 2: ODI's Energy Data Spectrum

The ODI Data Spectrum definitions are:

- **Open**: Data is made available for all to use, modify and distribute with no restrictions;
- **Public**: Data is made publicly available but with some restrictions on usage;
- **Shared**: Data is made available to a limited group of participants possibly with some restrictions on usage;
- **Closed**: Data is only available within a single organisation.

The followings sections describe how these can be applied across an organisation.

[7] https://es.catapult.org.uk/reports/energy-data-taskforce-report/

Open

This is data that has no restrictions on its access, whereby it is made available externally and there is no need or method to understand who has accessed the data. The use of Open data is to be governed by an Open Data Licence, based on the Government's Open Government Licence or other suitable licence.

Public

Data classified as Public is Open in terms of it not requiring any identification of who has accessed and is utilising the data, however, unlike Open data there are additional limitations on the user over above an Open Data Licence, namely that it cannot be shared in an adapted form.

Shared

Shared data uses the same licence as Public data, however, due to the nature of it a record of the people with access to the data is maintained. There is, by exception, reviews of the individuals or organisations accessing of Shared data, where for existing or developing reasons may have their access denied or revoked.

Closed

Data within this category is not shared publically as a matter of course. There may be instances where this data is made available using a bespoke Data Sharing Agreement (DSA) (or Non-Disclosure Agreement [NDA]) or as part of a wider contract with the organisation, where confidentiality clauses are already included.

Classifying the dataset

Utilising the classifications above, the data dictionary, metadata and answers to the data triage questions a specific data classification should be determined for the dataset. Figure 5 provides an example of the process flow and mitigation options to select the appropriate data classification.

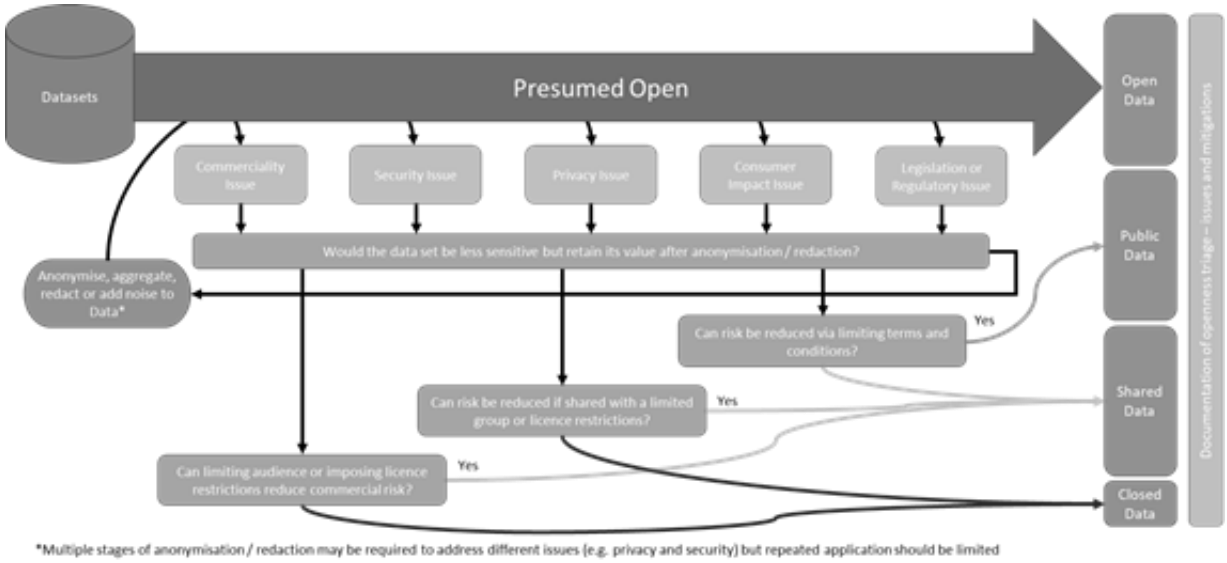

Figure 5: Example flow to select data classification

It is recommended to document a brief summary explaining the rationale and reasoning for the selected classification. An example data classification summary is shown in Figure 6.

Form Completeness		Complete	
		Initial Risk	Mitigation	Residual Risk
	Privacy	Red	Simple Data Manipulation	Amber
	Security	Amber	Simple Data Manipulation	Green
	Commercial	Green		Green
	Negative Consumer Impact	Green		Green
	Other	Green		Green
Summary	Green	86%		
	Amber	14%		
	Red	0%		
Openness Recommendation		Open		
Recommendation Notes		Limited residual risk remains, consider if areas of risk can be further mitigated through data modification. If this is not possible, consider if this is an acceptable level of risk to proceed with Open publication (subject to Data Protection Officer approval).		

Figure 6: Example triage and classification capture

Documentation

Following mitigation and data classification a finalised and consolidated set of data dictionary and metadata information should be captured. This information should be shared along with the specific dataset to inform the data user.

Where a dataset is classified as Closed, it is proposed that the metadata captured, as a minimum, is shared to provide future potential data requesters and data users' appropriate and relevant information.

The data triage documentation, responses to determined questions and mitigation techniques should be routinely captured with the summary shared to inform on the data classification selected.

Identify Release Mechanism

Based on the data requester, their intended use of the data and its defined data classification an appropriate release mechanism should be selected. Common options for release are likely to be:

- Direct email;
- Data Portal;
- Encrypted data exchange mechanism;
- Physical data transfer
- Username and password secured data site.

Direct email

The sharing of data directly via email should be avoided where another mechanism exists, such as an Data Portal where the dataset is attributed an appropriate data classification or an encrypted data exchange mechanism for Closed datasets as an example.
Where email is used the data, where required, should be password protected and the password shared in another email or via a different medium, such as an encrypted text messaging service.

Data Portal

Where the dataset is classification as open and possibly public every effort should be made to make the dataset available to all, beyond simply the data requester. This could be via an organisation's website but preferable a Data Portal, such as a CKAN instance.

Encrypted data exchange mechanism

Where it is not possible to share via a Data Portal, due to availability or classification of the data and encrypted data exchange mechanism should be used. A number of cloud data storage systems provide this availability such as MS Teams and Google Drive, where two-factor authenticated access can be provided for a defined period of time. There are also open source data exchange mechanisms available, such as 7-Zip .

Physical Data transfer

In some instances it may be required to provide data through a physical medium, USB, portable hard drive or other. Where this is necessary a secure postal service must be used, required a signature to an organisation or individual's registered address using a signature required service. A password should also be included, where it is required to access any and all data within the system, where the password is provided separately to the device.

Username and password secured data site

Where it is deemed that data should be made available to a range of people but those people can be tracked, most likely Shared data classification, it is proposed that a site requiring a unique username and password is utilised. This can be a Data Portal, where appropriate data classification and restrictions can be included.
Appropriate management of this system is required with review periods for people's access and regular and routine password changes required.

Sign Off and Review

Sign off by the relevant data role(s) within the organisation is required for the data release assessing the information identified in the documentation section as well as the dataset itself.

Who within an organisation should approve the dataset's release should be driven by the maturity of the data triage process in the organisation and risk level of the data. It is proposed that the approval levels within are used where the data triage process is mature. Where the Data Triage maturity is low the Data Manager and Data Owner should approve all datasets for release.

Table 13: Data approval release levels

Data Classification	Data Risk	Approver
Open	Green	Data Curator/ Operator
	Amber	Data Owner
Public	Green	Data Curator/ Operator
	Amber	Data Owner
Shared	Green	Data Owner
	Amber	Data Owner and Data Manager
Closed	Red	Data Manager

This data shall be centrally captured and used to define a routine review of the triage process and data classification. Table 14 demonstrates the review period based on data classification and data risk level.

Table 13: Data review periods

Data Classification	Data Risk	Review Period
Open	Green	Annual
	Amber	Bi-annually
Public	Green	Annual
	Amber	Bi-annually
Shared	Green	Bi-annually
	Amber	Quarterly
Closed	Red	Quarterly

Feedback

Feedback is key to ensure that appropriate and effective data is provided. There are a number of mechanisms to facilitate the provision of feedback. These include integrating direct feedback on datasets as part of a Data Portal, either in a structured manner or akin to commenting on a social media platform, a specific form to facilitate feedback to the organisation or ad hoc feedback via email.

It is proposed that a structured approach is taken to provide a level of commonality to facilitate an appropriate process for managing data feedback. It is suggested that the following elements could be utilised to facilitate a structured feedback mechanism:

- Name (person providing feedback)*:
- Contact email*:
- Organisation:
- Relevant dataset* (using title or identifier as part of dataset's metadata):
- Relevant field (using the tile as part of data dictionary):
- Feedback theme*:
    - Quality, Completeness, Structure, Inconsistency, Access, Definitions, Other
- Description for other* (where Other is selected above):
- Summary*:
- Criticality*:
    - Low - is not spotting the task being completed that the data supports
    - Medium – is affecting the quality or timeliness of the task being completed that the data supports
    - High – proposed task utilising the data cannot be completed

A service level agreement (SLA) should be agreed to provide a confirmation of receipt of feedback provided and a further SLA on providing confirmation back as to any data or process change based on the feedback as well as a timescale to provide the proposed revisions.

Challenge
Where there has been a request for data from an individual or organisation and that data request has not been facilitated due to an issue at any point throughout the playbook it is likely that a challenge may be received. A formal approach to managing this should be implemented, however, this will likely be dependent on an organisation's approach to managing similar challenges / complaints within other areas of that business.

Acknowledgements

Energy Systems Catapult

Various elements of this document have been taken and adapted from Energy Systems Catapult's Open Data Triage Implementation Guide.
Link: https://usmart.io/org/esc/discovery/discovery-view-detail/a350ef18-91fd-4ff6-8c0e-9ae22bc422f0
Licence: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

Western Power Distribution

Various elements of this document have been taken and adapted from Western Power Distribution's Presumed Open Data (POD) Network Innovation Allowance (NIA) project.
Link: http://www.westernpower.co.uk/projects/presumed-open-data-pod
Licence: http://www.westernpower.co.uk/terms-and-conditions

Energy Networks Association
4 More London Riverside
London SE1 2AU
t. +44 (0)20 7706 5100
w. energynetworks.org
@EnergyNetworks