Data Best Practice, Supporting Information

Publication date:	15 November 2021	**Contact:**	Greg Johnston, Senior Manager – Market Data Policy
		Team:	Modernising Energy Data
		Tel:	0141 354 5415
		Email:	ofgemdataservices@ofgem.gov.uk

This document provides supporting information to practitioners complying with Data Best Practice (DBP) guidance.

Document version	Description	Changes since previous document
Data Best Practice Supporting Information v0.3	Proposed Version May 2021	N/A
Data Best Practice Supporting Information v1.00	Version as a result of consultation May 2021	Minor updates to the Guidance element, as shown in. As set out in "Track Changes DBP Guidance v1.00"

Contents

1. Introduction

1.1. This document restates the principles and explanations in Data Best Practice Guidance and additionally provides supporting information in the form of techniques and examples to help practitioners as they comply with Data Best Practice Guidance.

1.2. The work of the other organisations, such as GO FAIR and Government Digital Service (GDS) strongly informed our development of the guidance and supporting information. The Government Digital Service (GDS) provides wide-ranging support for topics relating to data and digitalisation; it gives information and methods that span all the principles in the guidance. Of particular relevance are the following:

- GO FAIR and its FAIR data principles[1];
- GDS Service Standard[2];
- GDS Technology Code of Practice[3]; and the
- GDS Service Manual[4].

1.3. Where this supporting information refers to templates, find these in the zip file containing all templates hosted on ofgem.gov.uk[5]

[1] https://www.go-fair.org/fair-principles/
[2] https://www.gov.uk/service-manual/service-standard
[3] https://www.gov.uk/government/publications/technology-code-of-practice/technology-code-of-practice
[4] https://www.gov.uk/service-manual
[5] https://ofgem.gov.uk/publications-and-updates/data-best-practice-and-digitalisation-strategy-and-action-plan-consultation

General feedback

1.4. We believe that feedback is at the heart of good policy development. We are keen to receive your comments about this supporting information. We'd also like to get your answers to these questions:

- Do you have any comments about the overall quality of this supporting information?
- Was it easy to read and understand?
- Any further comments?

1.5. Please send any general feedback comments to ofgemdataservices@ofgem.gov.uk

2. Summary

2.1. DBP Guidance is designed to ensure data is treated as an asset and used effectively for the benefit of consumers, stakeholders and the Public Interest. It is a principles-based approach which provides guidance on the quality, accuracy and accessibility of data. It includes the principle that Data Assets must be treated as Presumed Open[6] which means that data must be made available for all people to use, unless the organisation responsible for handling the data provides specific evidence to show that the data should be withheld or its availability reduced (e.g. to protect individuals' rights to privacy). By complying with this guidance organisations will enable the full benefits of data to be unlocked for consumers.

Data Best Practice principles

1. Identify the roles of stakeholders of Data Assets
2. Use common terms within Data Assets, Metadata and supporting information
3. Describe data accurately using industry standard Metadata
4. Enable potential Data Users to understand Data Assets by providing supporting information
5. Make Data Assets discoverable for potential Data Users
6. Learn and deliver to the needs of current and prospective Data Users
7. Ensure data quality maintenance and improvement is prioritised by Data User needs
8. Ensure Data Assets are interoperable with Data Assets from other data and digital services
9. Protect Data Assets and systems in accordance with Security, Privacy and Resilience best practice
10. Store, archive and provide access to Data Assets in ways that ensure sustained benefits
11. Treat all Data Assets, their associated Metadata and Software Scripts used to process Data Assets as Presumed Open

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

Definitions

Data Asset: Any entity that is comprised of data. For example, a database is a data asset that is comprised of data records. A data asset may be a system or application output file, database, document, or web page. A data asset also includes a service that may be provided to access data from an application. For example, a service that returns individual records from a database would be a data asset. Similarly, a web site that returns data in response to specific queries (e.g., www.weather.com) would be a data asset.

This definition is taken from National institute of Standards and Technology (NIST).[7]

Data Contact Point: An organisation or individual who is the primary point of contact about a Data Asset or Metadata associated with a Data Asset.

Data Controller: A person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of a specific Data Asset.

This is based on the Information Comissioner's Office (ICO) definition but has been modified by removing reference to personal data and replacing it noting the processing of a Data Asset.[8]

Data Custodian: A person, public authority, agency or other body that has a legal right to process and publish a Data Asset as the Data Controller or otherwise.

Data Processor: A person, public authority, agency or other body which processes Data Assets on behalf of the Data Controller.

This is based on the ICO definition but has been modified by removing reference to personal data and replacing it noting Data Assets.[9]

[7] https://csrc.nist.gov/glossary/term/data_asset

[8] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/controllers-and-processors/what-are-controllers-and-processors/

[9] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/controllers-and-processors/what-are-controllers-and-processors/

Data Subject: The identified or identifiable living individual or entity to whom data relates.

Data User: An organisation or individual which utilises data held by a Data Custodian for any reason.

Data Best Practice Guidance: means (1) the guidance document issued by the Authority[10] in accordance with Part D of Special Condition 9.5 (Digitalisation) of the RIIO-2 price controls for Electricity Transmission, Gas Transmission and Gas Distribution and Special Condition 2.11 (Digitalisation) of the RIIO-2 price controls for the Electricity System Operator and (2) part of Ofgem’s standards for data and digitalisation.

Digitalisation Action Plan: an organisation’s plan to digitalise its Products and Services prepared and published in accordance with Part B of Special Condition 9.5 (Digitalisation) of the RIIO-2 price controls for Electricity Transmission, Gas Transmission and Gas Distribution and Part B of Special Condition 2.11 (Digitalisation) of the RIIO-2 price controls for the Electricity System Operator.

Digitalisation Strategy: the strategic approach taken by an organisation to digitalise its Products and Services and evidenced by the archive prepared and published by the Licensees in accordance with Part A of Special Condition 9.5 (Digitalisation) of the RIIO-2 price controls for Electricity Transmission, Gas Transmission and Gas Distribution and Part A of Special Condition 2.11 (Digitalisation) of the RIIO-2 price controls for the Electricity System Operator.

Digitalisation Strategy and Action Plan Guidance: means (1) the guidance document issued by the Authority in accordance with Part C of Special Condition 9.5 (Digitalisation) of the RIIO-2 price controls for Electricity Transmission, Gas Transmission and Gas Distribution and Special Condition 2.11 (Digitalisation) of the RIIO-2 price controls for the Electricity System Operator and (2) part of Ofgem’s standards for data and digitalisation.

DSAP: A combination of both Digitalisation Strategy and Digitalisation Action Plan.

Energy System Data: all Data Assets for which an entity is a Data Custodian as a consequence of it exercising its rights and obligations under a licence granted under section 6 (1) or (1A) of the Electricity Act 1989 or section 7, 7ZA, 7A or 7AB of the Gas Act 1986”.

[10] The terms “The Authority”, “we” and “us” are used interchangeably in this document

Metadata: a set of data that describes and gives information about other data.

Open Data: Data Assets, their associated Metadata and Software Scripts used to process Data Assets that are made available for anyone to use, modify and distribute without restrictions.

Open Data Triage: The process carried out by a Data Custodian to determine if there is any evidence of sensitivities associated with Data Assets, their associated Metadata and Software Scripts used to process Data Assets if they are used as Open Data. These sensitivities are limited to those that:

- relate to peoples' rights to personal privacy;
- security needs;
- obligations from legislation and/or regulation;
- commercial requirements that, if not protected, will have a negative impact on Products and Services for end-consumers; and
- would have a negative impact on the Public Interest.

Where any of the above sensitivities are identified, Open Data Triage will also include the determination of how the Data Custodian can mitigate any risk associated with them, while also making the Data Assets, their associated Metadata and Software Scripts used to process Data Assets as open to stakeholders as possible. The Data Custodian should consider both processing of and/or whether providing different levels of access by different types of stakeholders to the Data Assets, their associated Metadata and Software Scripts used to process Data Assets would help to mitigate any identified risk.

Presumed Open: The treatment of Data Assets, their associated Metadata and Software Scripts used to process Data Assets as Open Data, subject to Open Data Triage.

Products and Services: Anything that a party can offer to a market for attention, acquisition, use or consumption that could satisfy a need or want.

Public Interest: The welfare or well-being of the general public and society.

Single Provider Product or Service: A product or service among the Products and Services provided by a Data Custodian where no alternative option or provider is available to parties seeking to access that product or service.

Software Scripts: A code and its programming documentation; including information on how to execute that code, that enables users to read, capture, process, store or transmit a Data Asset or Metadata.

the Authority: means the Gas and Electricity Markets Authority that is established under section 1 of the Utilities Act 2000

3. The Data Best Practice Principles

1. Identify the roles of stakeholders of Data Assets

Explanation

3.1. The Licensees must identify the Data Assets that it is the Data Custodian of; for these, the Licensees must also identify any relevant Data Subjects, Data Controllers and Data Processors. The Licensees must keep this information in logs.

Techniques

3.2. The Information Commissioners Office (ICO) provide guidance regarding the identification of the roles of stakeholders. The ICO provide their guidance in the context of personal data. The definitions used for this Principle have been adapted from this but made to be applicable to Data Assets more generally.

3.3. As it has served as the origin for definitions used by Data Best Practice guidance, the ICO's own guidance can aid compliance with this Principle. For example, the below ICO quote helps with understanding of whether somebody or an organisation is the Data Controller or Data Processor of personal data. Removing the word "personal" from the ICO paragraph equally provides an approach to determining whether a Licensees is a Data Controller or Data Processor for the purposes of this Principle.

"To determine whether you are a controller or processor, you will need to consider your role and responsibilities in relation to your data processing activities. If you exercise overall control of the purpose and means of the processing of personal data – i.e., you decide what data to process and why – you are a controller. If you don't have any purpose of your own for processing the data and you only act on a client's instructions, you are likely to be a processor – even if you make some technical decisions about how you process the data." [11]

3.4. It is recommended Licensees appoint a specific senior leader to be responsible for the overall strategy, management and implementation of this Principle.

[11] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/controllers-and-processors/

3.5. The following documents are templates that could be used to help when logging information about stakeholders.

Templates

[Stakeholders.xlsx]
[Stakeholders.csv]

2. Use common terms within Data Assets, Metadata and supporting information

Explanation

3.6. Licensees must enable Data Users to search for and link Data Assets and associated Metadata to Data Assets and Metadata provided by other organisations. Licensees must label and describe Data Assets and Metadata using a taxonomy that is commonly recognised by practitioners who use the Metadata across the relevant subject matter domain.

Techniques

3.7. Implementing a standard referencing protocol enables organisations to more easily understand semantics associated with terms and data when they are used, allowing for convergence on a unified language where there is common ground. Participating with the implantation of an industry wide data catalogue featuring an agreed glossary and mechanism for feedback enables convergence to be accelerated and discoverability of data to be drastically improved beyond the capabilities of an organisation operating alone.

3.8. An obvious technique is to implement a common glossary of terms. However, a proliferation of glossaries risks each using different language and so not solving the basic challenge of shared semantics and clarity over terms. For the energy sector there are numerous glossaries, each trying to help needs for clarity over terms in particular segments of the market, without being coordinated as a whole, here are some examples:

- IEC Electropedia[12]
- IEA Energy Categories[13]
- Electricity Distribution Standard Licence Conditions[14]

[12] http://www.electropedia.org/iev/iev.nsf/welcome?openform
[13] https://ukerc.rl.ac.uk/categories.html
[14] https://www.ofgem.gov.uk/publications-and-updates/standard-conditions-electricity-distribution-licence

- Electricity Generation Standard Licence Conditions[15]
- Electricity Interconnector Standard Licence Conditions[16]
- Gas Interconnector Standard Licence Conditions[17]
- Gas Shipper Standard Licence Conditions[18]
- Gas Supplier Standard Licence Conditions[19]
- Electricity Ten Year Statement definitions[20]

3.9. Using the data domain as a perspective, the trend of proliferating glossaries can also be seen:

- Dublin Core glossary[21]
- Open Data Handbook glossary[22]
- Open Definition[23]
- W3c glossary[24]

3.10. These glossaries of terms are a valuable resources to use, both today, as parties share data and Metadata, but also as starting points for converging on a more unified language. Efforts to deliver greater unification of language and to standardise naming conventions used across a range of infrastructure domains are being undertaken by the Digital Framework Task Group[25] (DFTG) as part of the National Digital Twin programme of work. The long term goal of this Principle and the DFTG is to define an ontology that enables different sectors to use a common language which in turn better enables cross sector data sharing.

[15] https://www.ofgem.gov.uk/publications-and-updates/electricity-generation-licence-standard-conditions
[16] https://www.ofgem.gov.uk/publications-and-updates/electricity-interconnector-licence-standard-conditions
[17] https://www.ofgem.gov.uk/publications-and-updates/electricity-interconnector-licence-standard-conditions
[18] https://www.ofgem.gov.uk/publications-and-updates/gas-shipper-licence-standard-conditions
[19] https://www.ofgem.gov.uk/publications-and-updates/standard-conditions-gas-supply-licence
[20] https://www.nationalgrideso.com/document/157451/download
[21] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[22] http://opendatahandbook.org/glossary/en/
[23] https://opendefinition.org/
[24] https://www.w3.org/2003/glossary/subglossary/All/
[25] https://www.cdbb.cam.ac.uk/subject/dftg

3.11. To avoid an unnecessary proliferation of glossaries of terminology, we recommend a three staged approach:

1. Label data with keywords, citing the authoritative source of their definition i.e. Term [Glossary Reference] and for Example: "subject":"solar power station [http://www.electropedia.org/], energy cost [http://www.electropedia.org/]";
2. Adopting an authoritative glossary from existing sources instead of creating a new one, and then expanding on this only when necessary or in response to user feedback and challenge; and
3. Supporting the efforts of glossary providers and organisations like the DFTG, such as highlighting to them overlaps and inconsistencies between glossaries as you encounter them.

3. Describe data accurately using industry standard Metadata

Explanation

3.12. The Licensees must make it easy for Data Users to be able to use and understand information that describes each Data Asset. The Licensees must therefore provide Metadata associated with Data Assets and this Metadata must be made available to Data Users independent of the Data Asset.

3.13. The Licensees must treat the Metadata as a Data Asset. When providing Metadata, the Licensees must format and structure this in a widely recognised and accepted format that is machine readable.

3.14. There is no requirement for the Licensees to create Metadata about its Metadata associated with Data Assets.

3.15. When it updates or extends a Data Asset, the Licensees must ensure that the Metadata reflects any such changes so that Data Users can identify additions or changes.

Techniques

3.16. An example of a widely recognised method of formatting and structuring Metadata is through use of the Dublin Core[26] and in particular its 'Core Elements' Metadata standard (Dublin Core) ISO 15836-1:2017[27]. Dublin Core is a well-established standard for describing datasets and has many active users across a number of domains including energy sector users such as the UK Energy Research Centre (UKERC) Energy Data Centre[28]. It features a small number of key fields that provide a

[26] https://www.dublincore.org/
[27] https://www.iso.org/standard/71339.html
[28] https://ukerc.rl.ac.uk/

basic minimum level of description using established semantics, which can be built on and expanded as required.

3.17. There are 15 'core elements' as part of the Dublin Core standard which are described as follows:

Table 1: Dublin Core Standard

Element	**Description**
Title	Name given to the resource
Creator	Entity primarily responsible for making the resource
Subject	Topic of the resource *(e.g. Keywords from an agreed vocabulary)*
Description	Account of the resource
Publisher	Entity responsible for making the resource available
Contributor	Entity responsible for making contributions to the resource
Date	Point or period of time associated with an event in the lifecycle of the resource
Type	Nature or genre of the resource
Format	File format, physical medium, or dimensions of the resource
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)*
Source	Related resource from which the described resource is derived *(e.g. Source URI or DOI)*
Language	Language of the resource *(Selected language(s) from a agreed vocabulary e.g. ISO 639-2* [29] *or ISO 639-3* [30]*).*
Relation	Related Resource *(e.g. related item URI or DOI)*
Coverage	Spatial or temporal topic of the resource, spatial applicability of the resource, or jurisdiction under which the resource is relevant
Rights	Information about rights held in and over the resource

[29] https://www.iso.org/obp/ui/#iso:std:iso:639:-2:en

[30] https://www.iso.org/obp/ui/#iso:std:iso:639:-3:en

3.18. In the above table, the core descriptions are taken from the Dublin Core Metadata Initiative (DCMI)[31] licenced via Creative Commons (CC), CC BY 3.0 [32] - where edits or additions have been made by us, these are marked in *italics*.

3.19. Many of the elements required by Dublin Core are straight forward to populate either manually or through automated processes that ensure metadata is up to date and accurate. However, other fields are more open to interpretation. In table two we have listed best practice tips to help provide consistency across organisations.

Table 2: Best practice tips on Dublin Core

Element	Guidance
Title	This should be a short but descriptive name for the resource. • Be specific - generic titles will make useful resources harder to find e.g. 'Humidity and temperature readings for homes in Wales' is better than 'sensor data' • Be unique - avoid reusing existing titles where possible to help users find the required resources more effectively • Be concise - ideally less than 60 characters in order to optimise search engine display but many storage solutions will have an upper character limit (e.g. UKERC Energy Data Centre limit to 100 characters[33]) - use the other metadata fields for keywords and longer description. Avoid repetition • Title is distinct from a filename, which serves a more technical purpose - for titles: describe your resource in a way that any other Data User will understand e.g. use spaces between words, not underscores
Creator	Identify the creator(s) of the resource, individuals or organisations. • Creator - a creator is the primary entity that generated the resource being shared, this can be the same as the publisher, but isn't always • Unique Identifier - where possible use an authoritative, unique identifier for the individual or organisation e.g. company number as issued by Companies House[34]

[31] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[32] https://creativecommons.org/licenses/by/3.0/
[33] https://ukerc.rl.ac.uk/DC/cgi-bin/edc_submit.pl
[34] https://www.gov.uk/government/organisations/companies-house

Subject	Identify the key themes of the resource • Keywords - select keywords (or terms) that are directly related to the resource, try to avoid making assumptions about the reason a Data User is looking for data and limit the keywords used accordingly. • Glossary - select keywords from an agreed vocabulary e.g. UKERC Energy Data Centre uses the IEA energy balance definitions
Description	Provide a description of the resource that is inclusive and so can be read and understood by the full range of potential Data Users. • Overview - the description should start with a high level overview that enables readers to quickly be orientated about the context and content of the resource • Accessible - use language that is understandable to all, avoid jargon and acronyms. • Accuracy - ensure that the description objectively and precisely describes the resource without making implicit assumptions about its application, peer review can help identify potential issues. • Quality and limitations - include detail about known facets relating to the quality of the resource and any known limitations or issues. Comments on quality must be limited to what is objectively the case and not make presumptions about whether this quality is sufficient for any particular applications, as that would imply a presumption about how the resource will be used. • Core Supporting Information - ensure that any important supporting information is referenced in the description.
Publisher	Identify the organisation or individual responsible for publishing the data. This is typically the same as the metadata author. • Publisher - a publisher is the entity that is making the resource available, this can be the same as the creator or contributor • Unique Identifier - where possible use an authoritative, unique identifier for the individual or organisation e.g. company number as issued by Companies House[35]
Contributor	Identify the contributor(s) of the resource, individuals or organisations. • Contributor - a contributor is one or more entities that provided input to the resource which is being shared. These can be the same as the publisher • Unique Identifier - where possible use an authoritative, unique identifier for the individual or organisation e.g. company number as issued by Companies House

[35] https://www.gov.uk/government/organisations/companies-house

Date	Date is used for a variety of purposes (eg start of development/collection, end of development/collection, creation of resource, publication of resource, date range of data etc.). The usage of a date field should therefore be explained in the resource description and it may be necessary to expand beyond the Dublin Core elements and include multiple data fields. The nature of data will dictate the most valuable use of this field. For example, where data collection is concerned providing the time interval during which the data has been collected is likely to be most informative to Data Users, rather than the publication date. • Standardisation - the use of a standardised date format is recommended e.g. ISO 8601 timestamps[36] with explicit UTC offset. • Timezone - time values should always be given context including timezone and any modifiers (e.g. daylight savings). • Time Interval - it is recommended that time intervals use the ISO 8061[36] standard representation e.g. start/end, start/duration, etc.
Type	It is recommended to Identify the type of the resource using the DCMI type vocabulary[37]
Format	Identify the format of the resource - in the case of Data Assets, this is the file or encoding format. e.g. csv, JSON, API, etc.
Identifier	Provide a unique identifier for the resource • Identifiers - Digital Object Identifiers (DOIs) and Uniform Resource Locators (URIs) provide a truly unique identifier, but where these are not available then a system or organisation specific identifier can be of use.
Source	Identify the source(s) material of the derived resource • Identifiers - DOIs and URIs provide a truly unique identifier but where these are not available then a system or organisation specific identifier can be of use
Language	Identify the language of the resource • Vocabulary - Standard vocabularies such as ISO 639-2 [38] or ISO 639-3 [39] provide good ways to avoid ambiguity.

[36] https://www.iso.org/iso-8601-date-and-time-format.html

[37] https://www.dublincore.org/specifications/dublin-core/dcmi-type-vocabulary/2000-07-11/

[38] https://www.iso.org/obp/ui/#iso:std:iso:639:-2:en

[39] https://www.iso.org/obp/ui/#iso:std:iso:639:-3:en

Relation	Identify other resources related the resource. • Identifiers - DOIs and URIs provide a truly unique identifier but where these are not available then a system or organisation specific identifier can be of use. • Supporting Information - where additional resources are required to understand the resource these should be referenced.
Coverage	Identify the spatial or temporal remit of the resource. • Standardisation - utilising standard, authoritative spatial identifiers is recommended e.g. Unique Property Reference Number (UPRN), Unique Street Reference Number (USRN), Local Super Output Area (LSOA), Country Code, etc.
Rights	Specify which licence conditions the resource is controlled by. It should be clear if the resource is open (available to all with no restrictions), public (available to all with some conditions e.g. no commercial use), shared (available to a specific group possibly with conditions e.g. commercial data product) or closed (not available outside of the data custodian organisation). This field is vitally important as without clear articulation of user rights the resource cannot be used responsibly by Data Users. • Common Licences - where possible use standard licence terms, the following offer a range of standard licence conditions: ○ Creative Commons (CC)[40] ○ Open Government Licence (OGL)[41] ○ Open Data Commons (ODC)[42]

3.20. Metadata should be available independently of the Data Asset it is associated to and so stored in an independent file from the Data Asset, and this must be in a machine-readable format, such as JSON[43], YAML[44] or XML[45]. These machine-readable formats can easily have their data converted to be presented in a human readable format, such as using text editors. This approach ensures that Metadata can be shared independently from the Data Asset and that it is commonly accessible and not

[40] https://creativecommons.org/
[41] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
[42] https://www.opendatacommons.org/
[43] https://www.json.org/
[44] https://yaml.org/
[45] https://www.w3.org/XML/

restricted by software compatibility. The DCMI have provided schemas for representing Dublin Core in XML[46] and RDF[47], which may be of help.

3.21. Where a dataset is updated or extended the Data Custodian should ensure that the metadata reflects this such that potential users can easily identify the additions or changes. Where the data represents a new version of the dataset (i.e. a batch update or modification of existing data) then it is sensible to produce a new version of the dataset with a new metadata file. Where the dataset has been incrementally added to (e.g. updates to time series data) it may be best to update the metadata (e.g. data range) and retain the same dataset source.

Example

The Department for Business, Energy and Industrial Strategy (BEIS) publishes a dataset relating to the installed cost per kW of Solar PV[48] for installations which have been verified by the Microgeneration Certification Scheme,

```
{
"title":"Solar PV cost data"
"creator":"Department for Business, Energy and Industrial Strategy"
"subject":"solar power station [http://www.electropedia.org/], energy cost
[http://www.electropedia.org/]"
"description":"Experimental statistics. Dataset contains information on the cost per kW
of solar PV installed by month by financial year. Data is extracted from the
Microgeneration Certification Scheme - MCS Installation Database."
"publisher":"Department for Business, Energy and Industrial Strategy"
"contributor":"Microgeneration Certification Scheme"
"date":"30/05/2019"
"type":"Dataset"
"format":"XLSX"
"identifier":"19c3b55d-32dc-4bb3-8141-32bb2175affc"
"source":"https://certificate.microgenerationcertification.org/"
"language":"English"
"relation":""
"coverage":"Great Britain (GB)"
"rights":"Open Government Licence v3.0"
}
```

[46] https://www.dublincore.org/schemas/xmls/
[47] https://www.dublincore.org/schemas/rdfs/
[48] https://www.gov.uk/government/statistics/solar-pv-cost-data

3.22. Note, this example does not include a URI or DOI so a platform ID has been used as identifiers, this is not ideal as it may not be globally unique but provides Data Users with the best possible solution until a URI or DOI becomes available.

3.23. Note, the data.gov.uk[49] platform holds Metadata[50] for files in JSON format but the above has been simplified for the purposes of providing an example.

[49] http://data.gov.uk/

[50] https://ckan.publishing.service.gov.uk/api/action/package_search?q=%22Solar%20PV%20cost%20data%22

4. Enable potential Data Users to understand the Data Assets by providing supporting information

Explanation

3.24. Throughout the lifecycle of a Data Asset the Licensees must make available supporting information that Data Users require for the benefits to be gained by consumers, stakeholders and the Public Interest. The Licensees must ensure a point of contact is provided for Data Users to raise and resolve enquiries about the Data Asset and its supporting information.

Techniques

Core supporting information

3.25. When Data Assets are published openly, made publicly available or shared with a specific group it is critical that any core supporting information that is required to make the data useful for Data Users is also made available. It can be helpful to differentiate between core supporting information, without which the data could not be understood by even the most advanced and knowledgeable Data Users and additional supporting information, which makes opportunities to use Data Assets more inclusive to all potential Data Users.

3.26. A rule of thumb for core supporting information is that if the original Data Custodian of a Data Asset were to stop working with it for ten years and then start using it again (with the same level of domain expertise), information is available such that the Data Asset is intelligible and recoverable for use again as it had been the ten years prior.

3.27. The following topics for areas where core supporting information is typically required:

- Data structure description (e.g. data schema or relevant data models).
- Granularity (e.g. spatial, temporal, etc.) of the data. Including reference information, for example: with time-series data, it is specified whether data points represent the mean, start, midpoint or end value of the time intervals at which the data is reported.
- Units of measurement.
- Version number of any reference data that has been used
- References to raw source data (given within the Metadata)

- Processing methods (including the specific Software Scripts) that are applied to the data to carry out actions such as data validation, data transformation, data calculation.

3.28. It may be possible to reduce the effort required to provide core supporting information by using standard methodologies and data structures (e.g. ISO 8601[51] timestamps with UTC offset are strongly recommended) but note that there is a risk that an externally hosted reference dataset or document may not be enduring, unless it is archived and provided by an authoritative, sustainable body (e.g. ISO, British Standards Institution (BSI), UK Data Archive, UKERC Energy Data Centre, etc.). If there is doubt that a key reference data source or document will be available in perpetuity this should be archived by the publisher and, where possible, made available as supporting information.

Additional supporting information

3.29. The Licensees may be able to take advantage of opportunities to enhance the benefit Data Assets have to their own organisation. This might take the form of maximising engagement with Data Assets to solve a particular problem or drive innovation. Good examples of this how this can be achieved are data science competitions, which are commonly used as an open approach to solving specified problems, therefore drawing on the capabilities of a large and diverse group of interested stakeholders and subject matter experts. Additional supporting information is typically focused on descriptions of problem in non-technical language, common problems and how to identify them and opportunities for questions to be asked and answered.

3.30. The Licensees should also consider what additional supporting information it needs to make available in order to ensure inclusivity among Data Users. The exact nature of this information will be highly dependent on the Data Asset in question, and so there will be a strong need for engagement with Data Users to learn their needs with respect to additional supporting information. The techniques provided to support the DSAP Guidance give many methods relating to how to learn the needs of stakeholders and will be applicable here. The general objectives that can be met by providing supporting information are to:

[51] https://www.iso.org/iso-8601-date-and-time-format.html

- maximise user participation with Data Assets, for more efficient and effective use of these assets;
- reducing the number of subsequent queries about the Data Asset, allowing for greater autonomy and self-service by Data Users;
- helping share and re-use solutions associated with using a particular Data Asset; and
- highlighting particular challenges that the data publisher would like to drive innovators towards.

Example

Core Supporting Information

3.31. The UKERC are the hosts of the Energy Data Centre, which was set up to "to create a hub for information on all publicly funded energy research happening in the UK"[52]. The centre hosts and catalogues a large amount of data which is collected by or made available to aid energy researchers. Much of the data is made available using an open licence (e.g. Creative Commons Attribution 4.0 International License[53]). This licence enables data to be used for a wide range of purposes, including research. The centre aims to archive data for future researchers and as such, its normal activities embed many of the Data Best Practice into their work. This includes the provision of core supporting information that is required to understand the stored dataset. The custodians of the Energy Data Centre provided the '10 year' rule of thumb, described above.

3.32. The Energy Data Centre host the Local Authority Engagement in UK Energy Systems[54] data and associated reports. The data is accompanied by effective Metadata and a suit of core supporting information that enables users to understand the data. In this case, information includes:

- core supporting information
    - Description of the source datasets
    - Description of the fields and their units
    - ReadMe.txt files with a high level introduction to the associated project

[52] https://ukerc.rl.ac.uk/faq.html
[53] https://creativecommons.org/licenses/by/4.0/
[54] https://doi.org/10.5286/ukerc.edc.000005

- additional supporting information
    - A list of academic reports about the data and associated findings

3.33. The list of academic reports do also contain some of the core supporting information (e.g. the detailed collection methodology), but most of their content is additional supporting information.

3.34. Another example is the Kaggle VSB Power Line Fault Detection competition[55]. It asked participants to utilise sensor data to identify faults in power lines. To maximise engagement, the hosts provided high level overviews of the problem domain, more detailed explanations of datasets and offered additional advice as required through question-and-answer sessions that were widely published. Not all of this information was essential (ie core supporting information) for an expert to understand the data, but the underlying goal of the project was to attract new talent to the area and by providing additional supporting information, it was made easier to achieve this goal. If the potential participants could not readily understand the data and challenge they might have been more inclined to move on to another more lucrative competition.

Templates

3.35. The following template documents are provided to illustrate the type of field level information that should be provided to Data Users. Providing information in this format ensures information can be efficiently exported, owing to the use of machine-readable formats.

[Field Level Information.csv]
[Field Level Information.xlsx]

3.36. The documents below provide examples of the kind of structure that are typically useful to Data Users when providing supporting information.

[Supporting Information.csv]
[Supporting information.xlsx]

[55] https://www.kaggle.com/c/vsb-power-line-fault-detection/overview

5. Make Data Assets discoverable to potential Data Users

Explanation

3.37. Licensees must ensure that any potential Data Users can identify the Data Assets that the Licensees is the Data Custodian of, and how Data Users can pursue access to these Data Assets. Licensees must ensure that the Metadata associated to Data Assets is discoverable to Data Users, subject to the outcome of an Open Data Triage process.

Techniques

3.38. There is a distinction between a Data Asset being discoverable and it being accessible. In cases where a Data Asset is sensitive such that it can only be shared with a limited group of Data Users, it remains highly likely that the Metadata associated to this Data Asset can be made available without impinging on sensitivities. In particular where Data Users have no other alternative options for how to discover Data Assets (ie with respect to a Single Provider Product or Service) it is very hard to conceive of reasons why an Open Data Triage process would not find that Metadata is to be treated as open, except where acknowledging the existence of a Data Asset directly threatens national security.

3.39. The Data Best Practice Guidance describes Open Data Triage, whose purpose is to identify the appropriate level of accessibility for any given Data Asset, is relevant to both achieving the discoverability and accessibility of Data Assets (ie it is also applicable to Metadata).

3.40. For example, an advertising agency's Data Asset that describes personal details about individuals cannot be made openly available without the explicit consent of each Data Subject who feature in the Data Asset. However, advanced advertising products that provide customers with more effective Products and Services would be significantly less effective if the advertising agency could not explain to potential advertisers the types of data features they hold, as this helps to target advertisements to the right audience.

3.41. The following are techniques that can be used by Data Custodians to help potential Data Users find datasets.

Publishing Metadata

3.42. Metadata is effective at describing the contents and properties of a Data Asset. It is possible to make Metadata open (i.e. published with no access or usage restrictions) in most cases, and in the context of where a Data Asset can be described as a Single Provider Product or Service, its Metadata can be made open in all but the rarest of cases without creating security, privacy, commercial or consumer impact issues. Metadata can be published in common formats by individual organisations for re-use by Data Users (such as by Data Users external to the organisation whoa are creating a data catalogue of Metadata that spans across different organisations).

Webpage Markup

3.43. Where data is made available via services such as on a website, organisations can choose to markup datasets to make them visible to data-centric search engines and data harvesting tools. Structured markup has similarities to metadata but should not be seen as a replacement for standardised metadata.

3.44. For example, one of the most common uses of Schema.org[56] is to formalise the data that is held within online recipes. This is semi structured information that commonly includes a list of required ingredients, a list of equipment and step-by-step instructions. By structuring this data, cross-website recipe searches have been made possible. Services that have been derived from this include helpful tools, such as 'add to shopping basket' for food retailers who can integrate their sales with recipe websites.

Search Engine Optimisation

3.45. Search engines are likely to be the way many Data Users will discover Data Assets. The Data Custodian can therefore improve Data Asset discoverability by making sure data is presented in a way that search engines can find and index. Most major search engines provide guides explaining how to ensure that the correct pages and content appear in 'organic searches' (many search engines provide advice for webmasters

[56] http://schema.org/

e.g. Microsoft[57], Google[58]) and a range of organisations provide Search Engine Optimisation (SEO) services.

3.46. The Geospatial Commission[59] have developed a guide to help organisations optimise their websites to enable search engines to identify and surface geospatial data more effectively which also discusses SEO techniques.

Stakeholder Engagement

3.47. Direct stakeholder engagement is a focused technique which may range from physical events / webinars to bilateral meetings. Through this method the Data Custodian is able to target particular Data Users and provide specific advice and guidance to ensure their understanding of what Data Assets could be used by the Data User and which data the Data Custodian could be collecting / making more available.

Point of Contact

3.48. It is not practical to anticipate all of the needs Data Users will have with respect to the discoverability, by providing a clear point of contact, the Data Custodian can offer a better quality of service to Data Users.

Example

3.49. BEIS has published a dataset relating to the installed cost per kW of Solar PV for installations which have been verified by the Microgeneration Certification Scheme. This dataset is published under an Open Government Licence so the data can be accessed by all. The BEIS, as the Data Custodian, has registered the dataset with the UK Government open data portal Data.gov.uk[60], this makes the Metadata publicly available (albeit only via the API) in JSON format, which is a machine-readable data file that can easily be translated into a human-readable form. The way the Metadata has been shared additionally provides supports search engine optimisation, allowing the existence of the Data Asset to be returned in organic search; it also takes

[57] https://www.bing.com/webmaster/help
[58] https://www.google.com/intl/en_uk/search/howsearchworks/
[59] https://www.gov.uk/government/organisations/geospatial-commission
[60] https://data.gov.uk/

advantage of webpage markup to make the data visible to dataset specific search engines.

3.50. The discoverability actions taken for the above example are related to a dataset which is publicly available, but they are equally applicable to Data Assets that are not open and could be used for a metadata stub entry in a data catalogue.

6. Learn and deliver to the needs of current and prospective Data Users

Explanation

3.51. The Licensees must identify the Product and Service requirements of Data Users who use, or who wish to use, the Data Assets provided by them as Data Custodian. The Licensees must then develop and deliver Products and Services to meet Data Users' requirements, where doing so is of reasonable cost and would benefit at least one of the following: end-consumers, stakeholders or the Public Interest. Where the Licensees is obligated to publish and update a DSAP, these Products and Services must be integrated into those publications.

Techniques

3.52. Digital connectivity and the application of Data Assets enable a wealth of Products and Services across the economy and can create new Data Users outside of traditional sector silos. To maximise the value of data it is vital that Data Custodians develop a deep understanding of the spectrum of Data Users and their differing needs such that datasets can be designed to realise the maximum value for customers.

3.53. This requirement to understand Data Users' needs and so determine Data Users' requirements from Products and Services has overlap with the DSAP Guidance. Where an entity is required to comply with both the Data Best Practice Guidance and the DSAP Guidance, Data Users' needs from Data Assets are to feature in the DSAP.

3.54. The Techniques and Examples provided, below, provide information about deeper-dives into more technical methods than those described in the DSAP Guidance, which may be appropriate for the detailed development and design of Products and Services.

3.55. The table below provides examples of learnings that an organisation might need to gain from Data Users.

Table 3: Learnings that might be gained from Data Users

Topic	Description
Current and potential Data Users	Who is using your Data Asset? Who would like to use your Data Asset? Who should be using your Data Asset?
The goals of each current and potential Data User	What outcomes do Data Users aim to achieve by using your Data Assets?
The needs of each current and potential Data Users	Data Users will have a range of needs driven by many factors, such as differing objectives, systems and capability. Characterising these helps solutions to be developed. Technical needs are likely to include requirements relating to: • Data granularity (time, space, subject); • Data accuracy or precision (how closely does the data reflect reality); • Data timeliness and consistency (duration between data creation and access); • Functionality and simplicity of access (file download, API requests, etc); • Service: ○ Reliability (service availability over time); ○ Stability (consistency over time); ○ Agility (the ability to adapt to changing needs); and • Data interoperability (joining to other datasets).
How this relates to the goals of the Data Custodian	How do the objectives of the Data Users compare to and drive change to the objectives of the Data Custodian?
How this delivers benefit for end-consumers and the Public Interest	• Does meeting the need of the Data User provide benefits to end-consumers and/or the Public Interest? • If there are conflicting objectives between the Data Custodian and the Data User, how can these be overcome and is one providing significantly more benefits to consumers / the Public Interest?
How the Data Custodian can deliver Products and Services in an appropriate format within reasonable timescales	• Will Data Users benefit from some data being made available early, regardless of its current data quality? • Are there ways to incrementally improve the quality of providing data discoverability/access/data quality, rather than to make improvements for Data Users late on only after one large effort? • In what is it best to deliver Products and Services (eg are Data Users in need of data discoverability, or do they have a good understanding of what Data Assets they need access to, with the need being for that access to be gained).

3.56. To help gain clarity from Data Users about their needs[61], it can be beneficial to characterise their needs in ways, such as the following:

Table 4: Characterising Data User needs

Type of need	Comment
Explicit	How Data Users describe what they are trying to do and how they need to do it.
Implicit	Needs that are not directly expressed by Data Users and that at times they may not be aware of, but that are evident as being required to meet the objectives Data Users have described.
Created	Needs the Data User has that emerge due to ensuring benefits are gained by end-consumers' and the Public Interest.
High-level	For example: 'I need to understand the data so that I don't use it incorrectly'
Detailed	For example: 'I need to know about all data processing that has already been applied to the data so that when I measure the frequency of data collection failures, my calculations are precise to within three decimal places'.
Constraints	For example: 'I need to trust the data so I can defend my decision'

3.57. It is important to be aware that it is not feasible to identify all potential Data Users or applications before Data Assets are made available to Data Users. Therefore, data without a clear Data User or application should not be ignored and certainly needs to be discoverable as this will raise awareness about the Data Asset and enable both people and machines to innovate applications that take advantage of it.

Example

Structured requirements gathering

3.58. One approach that can be used to gain input from potential users is to discuss with Data Users from different backgrounds can the challenges they face and their needs. Formulating the needs of Data Users as structured 'user stories' can help generate a backlog of Product and Service requirements that help guide Product and Service development and delivery. A simple example of this kind of approach is to structure

[61] https://userresearch.blog.gov.uk/2017/12/05/understanding-the-needs-of-service-data-users/

needs in terms of: 'As a **[Role]** given that **[Situation]** I need **[Requirement]** so that **[Outcome]**', where the emboldened terms in brackets are:

Table 5: Structured requirements gathering

Term	Description
Role	The type of stakeholder
Situation	The motivating or complicating factor
Requirement	The user need
Outcome	The desired outcome

For example:

3.59. As a **homeowner** given that **home heating accounts for 31% of carbon emissions,** I need **to have access to data that evidences the impact of energy efficiency measures** so that **I can prioritise upgrades and developments and reach net zero carbon as quickly as possible.**

Templates

3.60. The following templates are examples of how user research describe in this section might be documented.

[User Research Documentation.odt]
[User Research Documentation.docx]

7. Ensure data quality maintenance and improvement is prioritised by Data User needs

Explanation

3.61. Where the Licensees reasonably expects the Data Users' application of Data Assets for which it is the Data Custodian to deliver a net benefit for end-consumers, stakeholders and/or the Public Interest, the Licensees must ensure that Data Assets are of a quality that is sufficient to meet reasonable requirements of its Data Users. Data Users must have an option for contesting decisions regarding the definition of sufficient data quality of a Data Asset.

3.62. Where data quality issues are identified the Licensees must ensure that these issues are logged, considered and rectified as soon as practicable.

Techniques

3.63. Data quality is subjective, and Data Assets do not have a universal data quality characteristic. The quality of data is dependent on the specific application it is to be used for. A Data Asset may have characteristics that are sufficient or even more than sufficient for a particular application, but that same Data Asset's characteristics may render it inadequate for another purpose. Therefore, fundamentally the required data quality of a Data Asset is not absolute and is instead and is instead defined by the applications it is expected to be used for now or in the future.

3.64. There are objective measures available for characterising data. Data accuracy (such as uncertainty associated with measurements), is an example of how data accuracy can be characterised, but data accuracy does not directly translate to data quality. Other examples of objective methods for characterising Data Assets include: reporting on the granularity of information; the extent to which sample data represents an overall population; and information about the basis for assumptions used when creating the data.

3.65. In many cases, Data Assets store information that can objectively be categorised as right or wrong e.g. customer addresses, asset serial numbers, etc. In these cases, the Data Custodian should seek to ensure its data collection, storage and sharing reflect accurate data and to conduct ongoing assurance to ensure accuracy is maintained and errors are identified and resolved.

3.66. Data Users' needs will implicitly and explicitly create data quality requirements that can be described in the form of objective characteristics of Data Assets. The most advanced requirements from Data Users that, if met, deliver benefits to consumers and the Public Interest, will define what constitutes "sufficient" data quality drive for any given Data Asset. This definition of "sufficient" quality will change over time, as Data Users needs evolve.

3.67. As well as improving the quality of Data Assets to be sufficient for Data Users' needs, Data Custodians must ensure they maintain a sufficient level of data quality and that they incrementally improve the quality of data such that it can be effectively utilised by its Data Users and so that the benefits gained by consumers and the Public Interest are maximised.

3.68. Given the breadth and depth of Data Assets involved in modern markets, such as the energy sector, and the fact that data quality needs are continually evolving, it is likely that at times the quality of Data Assets may not be sufficient for Data Users' needs. In these cases, organisations should not view data quality as a reason to delaying the making visible and opening of Data Assets. These Data Assets are likely to meet some of the needs of Data Users (or even all of the needs of some Data Users) and by providing Data Users the opportunity to work with these Data Assets directly, the process of improving data quality will be made easier, as Data Users will be able to provide direct feedback about their residual needs that are still unmet.

3.69. Where data quality is substantially inadequate at meeting Data Users' needs, the Data Custodian may need to consider strategic approaches to rectifying this, such as the addition of Products and Services or actions into the DSAP. Examples of the kinds of ways a Data Custodian might respond to this challenge are listed in table six, below:

Table 6: Data Custodian example approaches to improving data quality

Activity	Example
Changes to the underlying creation and collection of data at source	Such as new types of hardware (monitors and sensors) that create higher precision data than existing equipment
Changes to how data processing takes place	The way in which a data pipeline is designed, such as how it conducts data validation, preparation or aggregation.
Data retention and management practices	Changing practices to increase/decrease the deleting of data due to storage constants
Master Data Management (MDM)[62]	Which is a large topic, but can cover topics such as inputs, monitoring for consistency across systems and enabling the rectification of issues quickly when they are identified.
Updates to guidance information	Including into Metadata known issues where data quality issues are known to exist, thus enabling Data Users to self-manage these short-comings for the period while the limitations are managed and resolved.
Data science methods	In some cases poor quality data can be improved through the application of data science techniques (eg the application of Machine Learning to predict where specifically Data Asset issues are likely to occur). Even simple statistical methods, such as interpolation, can play a role. In addition, the act of opening up data and the processing methods applied to it can enable third parties to help the Data Custodian to rectify underlying data quality issues.
Liability Wavers	Where an organisation is aware it is sharing data tat has data quality issues, it may be prudent to use a liability waver to provide certainty to Data Users about what information can be trusted and what information is in need of remediation (though note, liability wavers do not remove all legal responsibility for data quality).

Examples

Accuracy

3.70. An organisation collects and holds data about companies that operate in a sector to help potential innovators find suppliers or collaboration partners. The data about this is made available on a website and can be queried by Data Users. A Data User searches

[62] https://www.sciencedirect.com/topics/computer-science/master-data-management

for an organisation and finds that a company has been incorrectly categorised. Because the data Custodian has provided a contact form, this enables the Data User to submit information about the need for a correction, which is subsequently verified by the Data Custodian and the dataset is updated.

Quality

3.71. A Data User wishes to develop an application that enables public transport users to select their preferred means of transport based on impact on air pollution in a city. The public transport provider makes a range of data available about the various modes of transport including routes travelled, vehicle type, average emissions and passenger numbers. However, the Data User has identified that the emissions data is not of a sufficient quality for their application. Table seven below gives examples of the types of issues a Data User might encounter and solutions the public transport provider could offer:

Table 7: Issues a data user might encounter

Data User need	Public transport provide example solution
The definition of "average" is not provided	Publish the definition of "average" used for the calculations. Or, better yet, publish the actual software script that holds the explicit instructions for how this calculation has been performed
The average emissions data field is not consistently populated	Use data engineering methods to create software that checks the dataset for missing values and populates these based on interpreted values using characteristics such as vehicle type information.
To understand the variation between the average emission output and the actual emissions in each local area.	Deploy a real emission monitoring solution and so enable reporting of a more granular dataset that provides information about each locality. Or Joining the data with an existing alternative Data Asset (eg static air quality monitoring sites) and processing this data as 'proxy' data from which local emissions data can be estimated.

8. Ensure Data Assets are interoperable with Data Assets from other data and digital services

Explanation

3.72. Licensees must enable interoperability, between the Data Assets for which they are Data Custodian and Data Assets of other Licensees as a minimum standard.

3.73. When the Licensees makes Data Assets available, it must do so in ways that make it reasonably easy for Data Users to gain information and/or insight from those Data Assets in conjunction with Data Assets from other Licensees. There must also be sufficient information to align to Data Assets from other industries.

3.74. The Licensees must make data available in such a way that it is reasonably easy for Data Users to:

(i). exchange Data Assets between systems;
(ii). interface with Data Assets held in the Licensees systems; and
(iii). join Data Assets with other Data Assets, such as by using standard interfaces, standard data structures and/or common reference data.

Techniques

3.75. The potential utility of data is increased when it can be shared, linked and combined with other data with ease and data interoperability describes the properties and aspects about data that help it to, for example, be shared and ported with ease between different systems, organisations and individuals and so contribute to the goal of ensuring ease of use of data between systems.

3.76. Data interoperability, however, requires bilateral participation by two or more parties, where each party manages and makes available is data in ways that minimise the effort associated with carrying out work that involves Data Assets from all parties. It is therefore beyond the control of a single party to guarantee the interoperability of data, as this depends on the decisions and actions of an independently governed other party. However, a single party can *enable* data interoperability. Where all parties are successfully enabling the interoperability of data, interoperability is achieved.

3.77. To enable interoperability, particularly where this must be achieved with multiple other parties and with parties for which there is not be a direct relationship, dialogue or communication, it is efficient to take advantage of recognised standards and widely used common practices. Where all parties are taking advantage of recognised standards and common practices, the residual effort to interoperate Data Assets provided by these parties is minimised. The residual effort is simplified to tasks such as translating data from one recognised standard to another, which is a task that readily lends itself to standardisation automation and scalability. Where parties have chosen to adopt the same standards/practices, even this will not be necessary.

3.78. Data interoperability is an expansive topic. It includes many technical facets relating to data, for example, how Metadata structured and formatted and the semantics of the language used to describe the data. This guidance already includes Principles dedicated to these topics. For this Principle, information is provided about topics not otherwise addressed.

Open Data Licenses

3.79. Where organisations publish Open Data or Software scripts – an associated Open Data Licence is usually specified which by its nature explains the restrictions placed upon the use of that Data Asset. The Open Data Institute (ODI) provides a useful explanation of the use of Open Data Licenses[63] By ensuring that each Data Asset has an associated Open Data Licence (See the Dublin Core explainer earlier in this document for how to implement this into the Metadata) organisations who wish to use the data will understand the scope through which they can access it.

3.80. Some common Open Data Licenses include:

- Open Data Commons: http://opendatacommons.org/licenses/pddl/
- Open Government Licence: http://www.nationalarchives.gov.uk/doc/open-government-licence/

[63] https://theodi.org/article/publishers-guide-to-open-data-licensing/

Standard Data Structures

3.81. Data structure standardisation is a common method of enabling the alignment of data across organisations and enabling seamless portability of data between systems. Standardisation enables robust interoperability between systems and if the standard has been correctly adhered to, it enables entire data structures to be ported between systems. However, standardisation of data structures can be time consuming and if not governed effectively, runs the risk of expensive refactoring of effort. Further, many standards already exist, each serving different purposes and different domains will feature differing levels of maturity with respect to the development of standards. Table eight below gives examples of where data structure standards have been implemented.

Table eight: Examples of Data structure standards implementation

Sector	Comment
Telecoms	Many of the underlying functions in telecoms have been standardised at a data structure level by the Global System for Mobile Communications[64] (GSM) Association (GSMA). This enables network companies to deploy devices from different manufacturers with limited burden, this drives down costs by reducing the risk of vendor lock in.
Energy	The Common Information Model[65] (CIM) is a set of standardised data models which can be used to represent electricity networks, their assets and functionality. CIM is being deployed in a number of network areas to aid the portability between systems, enable innovation and lower costs.

Standard Interfaces

3.82. Data interfaces can be standardised, this means that formal channels of communication are structured in a standard way that enables systems to 'talk' to one another with ease. This approach has the advantage of being quick to implement as required by Data Users, through the organisation providing the data giving robust documentation for its defined 'standard' approach to interfacing with it. This enables integration and Data Users and does so across a sector and sectors, with it being made easy to develop solutions for data integration.

3.83. However, this approach is limited in that a new interface needs to be developed or deployed for each type of data that needs to be shared. Additionally, in sectors where there a few powerful actors they can use interface standardisation to create siloed ecosystems which reduces portability. Below are examples of organisations offering Application Programme Interfaces (API) documentation that can offer a useful information for when interfacing with a system.

[64] https://www.gsma.com/

[65] https://www.iec.ch/smartgrid/standards/

Table nine: Examples of API documentation

Examples of API documentation	Comment
Google[66]	For exchanging data created by Google's Google Assistant
Amazon[67]	For exchanging data created by Amazon's Alexa
Apple[68]	Apple's HomeKit, for interfacing with Apple products
Hive[69]	For integrating with Hive products

Reference Data

3.84. Reference data is a subdomain of Master Data Management. Reference data can play many specific roles, but in general it is used to help 'match and link' how different datasets relate to one another or how a single dataset relates to concepts that are not included in the dataset itself.

3.85. Certain data plays a very important role as reference data, these important reference data are typically called 'data spines', the UK national address database plays an important role for the Office for National Statistics (ONS) as it carries out the UK census, carrying out the census effectively strongly relies on matching and linking datasets.

3.86. Matching datasets back to reference data spines can be a useful method to enable non standardised data to be joined with relative ease, avoiding the often more expensive need to change Data Assets at source. This approach can provide a basic level of interoperability across datasets without the need for full standardisation. However, it does require the users to learn about each new dataset rather than being able to understand the data from the outset with standard data structures.

3.87. When processing data from disparate sources it may not be directly possible to match fields of data within the data sets, they may not have common features. To tackle this,

[66] https://developers.google.com/assistant/sdk/reference/rpc
[67] https://developer.amazon.com/en-US/docs/alexa/alexa-voice-service/alexa-discovery.html
[68] https://developer.apple.com/homekit/
[69] https://developers.hive.com/reference

some data intensive organisations relate these data to major data spines, these major data spines can be about very broad subject matter domains, such as company reference number, individual identifiers and property identifiers. Then, as part of data ingestion pipelines, these organisations match datasets to one or more of these major data spines, which have a good chance of being successful at finding matches. This means that even where datasets do not seemingly contain related data, it can be possible to relate them and then go on to join and make use of them. This method also has the benefit of providing data spines that multiple organisations can more readily agree to work in coordination around and that it provides a low effort approach to a single organisation establishing its enterprise data architecture.

Example

3.88. Electricity network data is essential for a number of emerging energy system innovations including the successful integration of a highly distributed, renewables dominated grid. However, the network is divided into a number of areas which are operated by different organisations that have implemented different data structures to manage their network data (power flow model, GIS and asset inventory).

3.89. The CIM is the common name for a series of IEC standards (see the series IEC 61970 and IEC 61968[70]), these guide the standardisation of the data structure for electricity network data. The deployment of the CIM standards enables network operators to provide third parties with access to their data in a standard form that enables innovation to be rolled out across network areas with relative ease.

[70] https://www.iec.ch/smartgrid/standards/

9. Protect Data Assets and systems in accordance with Security, Privacy and Resilience (SPaR) best practice

Explanation

3.90. The Licensees must ensure that compliance with this guidance does not negatively impact its compliance with relevant regulations, legislation and SPaR requirements.

Techniques

3.91. The Data Best Practice Guidance is positioned in service of wider requirements, such as legislation and other regulation, that create requirements relating to SPaR. However, these other requirements are not in conflict with the goals of this guidance, which is designed to encourage organisations to adopt a risk-based approach to the delivery of its Products and Services and therefore to ensure that all factors that contribute to creating benefits to end-consumers and the Public Interest are considered when using data, and not only their need for SPaR.

3.92. The use of Agile approaches that seek to balance risk and reward are an effective means for striking a balance between SPaR and other needs in the context of a data (and digital) domain. The Ofgem Supporting Information to the DSAP Guidance contains advice about this approach to working.

3.93. Specifically, with respect to SPaR requirements, a number of frameworks, standards and regulations exist that provide organisations with implementable guidance, such as:

Table 10: SPaR requirements guidance

Guidance	Comment
National Cyber Security Centre (NCSC)'s Cyber Assessment Framework (CAF)[71]	CAF provides guidance for organisations responsible for vitally important services and activities.
Examples of standards	• Information Security Management, ISO 27000 [72] • Industrial communication networks, IEC 62443 [73] • Substation Automation, Protection, and Control Systems, IEEE C37.240-2014 [74] • Smart Cities. Specification for establishing and implementing a security-minded approach, PAS185 [75]
Examples of regulation	• Network Information Systems Directive[76] (NISD) • General Data Protection Regulation[77] (GDPR)

3.94. A wealth of advice can also be gained from centres of excellence, such as the following types of organisations:

- BEIS and Ofgem - The joint competent authorities[78] for downstream gas and electricity
- Ofgem's RIIO-2 Cyber Resilience Guidelines[79]
- National Cyber Security Centre[80] (NCSC)
- Centre for the Protection of National Infrastructure[81] (CPNI)
- International Electrotechnical Commission[82] (IEC)
- Institute of Electrical and Electronics Engineers[83] (IEEE)
- International Organization for Standardization[84] (ISO)

[71] https://www.ncsc.gov.uk/collection/caf/cyber-assessment-framework
[72] https://www.iso.org/isoiec-27001-information-security.html
[73] https://webstore.iec.ch/publication/7029
[74] http://standards.ieee.org/findstds/standard/C37.240-2014.html
[75] https://shop.bsigroup.com/forms/PASs/PAS-185/
[76] https://ec.europa.eu/digital-single-market/en/network-and-information-security-nis-directive
[77] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/
[78] https://www.ofgem.gov.uk/ofgem-publications/144069
[79] https://www.ofgem.gov.uk/publications-and-updates/riio-2-cyber-resilience-guidelines
[80] https://www.ncsc.gov.uk/
[81] https://www.cpni.gov.uk/
[82] https://www.iec.ch/
[83] https://www.ieee.org/
[84] https://www.iso.org/

- British Standards Institution[85] (BSI)

[85] https://www.bsigroup.com/

10. Store, archive and provide access to Data Assets in ways that ensures sustained benefits

Explanation

3.95. When Data Assets are not required by the Licensees, the Licensees must ask stakeholders whether they consider that the Data Assets could create a future benefit if archived. The Licensees must archive Data Assets when, taking account of stakeholders' views it determines that the storage will be a net benefit to consumers, stakeholders and/or the Public Interest.

3.96. When archiving, the Licensees must also ask stakeholders for views on the storage method and formats to use. In determining what to archive, it must consider:

- Data Assets;
- Metadata;
- Software Scripts used to process Data Assets;
- data derived resulting from this processing of the original Data Asset; and
- human-readable representations of the data and any other relevant supporting information.

3.97. The Licensees must ensure that the risk of unintentional or malicious deletion of Data Assets, Metadata and Software Scripts used to process Data Assets is effectively managed and monitored to ensure possible recovery.

Techniques

3.98. Technical storage solutions that offer component and geographical redundancy are common place offerings by cloud providers, who offer data resilience as standard86. However, Data Custodians still have responsibilities to ensure that the system is correctly configured and assessed to ensure that the desired system resilience and system security are achieved, as required.

[86] https://docs.aws.amazon.com/AmazonS3/latest/dev/DataDurability.html

3.99. Where practical, it is prudent to store the most granular version of Data Assets, as this preserves the opportunity to use the Data Asset for the widest range of future applications in future. However, there may be cases where the raw data is too large in scale to store it all indefinitely. In these cases, the Data Custodian will have to consider solutions that reduce the scale of the data, but preserve as much of its utility as practical (for example by applying aggregation methods or limiting the duration of retention of the most granular data).

3.100. Stakeholders ability to create benefits from data are not limited only to their ability to access the data, but also through their access to information about how the data has been used, how it is (or was) processed when it was in use and ways that the data was processed and presented when decisions were made based on it. When archiving data, the exact additional needs of future stakeholders will have some uncertainty, but this uncertainty can be reduced through the Data Custodian working openly with today's stakeholders and so improve decisions about which information to archive along with the Data Asset itself.

3.101. Where data is not or is no longer required by the original Data Custodian, it can be most beneficial to archive the data using a trusted third party. For example the UKERC Energy Data Centre[87], UK Data Archive[88]. Access to data should also be provide in a way that is appropriate for the status of the data. Table 11 below shows how live and historic data might be treated differently to each other.

[87] https://ukerc.rl.ac.uk/

[88] https://www.data-archive.ac.uk/

Table 11: Live and historic data

Data Type	Access Method
Live Data	Today, Application Protocol Interfaces (APIs) are the defacto standard for the delivery of live data. Many organisations from Twitter[89] to Elexon[90] have developed and deployed successful APIs which respond to the needs of their Data Users.
Historic Data	Machine readable files (such as CSV or JSON) are preferable for bulk historic data. This negates the need for a large number of API calls which reduces the load on the send and receivers' systems and provides the end user with a useful dataset which can be manipulated as required. The BEIS National Energy Efficiency Data[91] is a good example of providing well documented bulk data. The Met Office also publish a variety of useful data in bulk formats[92]. For the benefit of reducing the cost of long-term storage solutions, it may be appropriate for archived (often but not always historic) data that is accessed infrequently to have service level agreements where the time to wait for gaining access to the Data Asset is longer.

Example

3.102. The UK Smart Meter Implementation Program[93] is responsible for rolling out digitally connected meter points to all UK premises with the goal of providing an efficient, accurate means of measuring consumption alongside a range of other technical metrics. Electricity distribution networks can request to gain access to this data in order to inform their planning and optimise operation. The personal nature of smart meter data means that organisations have to protect individual consumer privacy and are therefore choosing to aggregate the data at feeder level before it is stored, this

[89] https://developer.twitter.com/en/docs
[90] https://www.elexon.co.uk/guidance-note/bmrs-api-data-push-user-guide/
[91] https://www.gov.uk/government/statistics/national-energy-efficiency-data-framework-need-anonymised-data-2019
[92] https://registry.opendata.aws/uk-met-office/
[93] https://www.gov.uk/government/publications/smart-metering-implementation-programme-information-leaflet

approach provides the network with actionable insight for the current configuration of the network.

3.103. However, network structure is not immutable. As demand patterns change and constraints appear, network operators may need to upgrade or reconfigure their network to mitigate problems. The decision to aggregate the data means that it is not possible to use the granular data to simulate the impact of splitting the feeder in different ways to more effectively balance demand across the new network structure. In addition, it means that the historic data cannot be used for modelling and forecasting going forwards. Finally, the data ingest and aggregation processes need to be updated to ensure that any future data is of value.

3.104. Northern Power Grid[94] (a GB Electricity Distribution Network) have proposed to store the smart meter data that they collect in a non-aggregated format but strictly enforce that data can only be extracted and viewed in an aggregated format. This approach is novel in that it protects the privacy of the consumer whilst retaining the flexibility and value of the data.

[94] https://www.northernpowergrid.com/

11. Treat all Data Assets, their associated Metadata and Software Scripts used to process Data Assets as Presumed Open

Explanation

3.105. The Licensees must treat all Data Assets, their associated Metadata and Software Scripts used to process Data Assets where it is the Data Custodian, as Presumed Open and these must be subjected to Open Data Triage.

3.106. The Licensees must treat information created during Open Data Triage as Open Data, except where this will result in a sensitivity listed in the Open Data Triage definition.

3.107. Where a sensitivity is identified with the Data Assets, their associated Metadata and Software Scripts used to process Data Assets, the Licensees must take all reasonable steps to provide suitable options to make them available in a format or version that mitigates the risk associated with any identified sensitivity. When identifying those options the Licensees should additionally consider whether providing different stakeholders with different levels of access would mitigate any identified risk while minimising any reduction in the utility of the Data Asset.

3.108. The Licensees must make available the Data Assets, their associated Metadata and Software Scripts used to process Data Assets in the changed formats, versions or with the different levels of access to stakeholders, where it is beneficial to end consumers, stakeholders and the Public Interest to do so.

3.109. The Licensees must record at least the following information about Open Data Triage processes:

- what has been triaged;
- when the process took place;
- a description of the sensitivities and risks, if any, that have been identified including the type of sensitivity as defined by Open Data Triage;
- the options considered for how to mitigate any sensitivities or risks identified and the impact these have on the utility of the Data Assets, their associated Metadata and/or Software Scripts used to process Data Assets; and
- any decisions made.

3.110. The Licensees must ensure there is a point of contact available to stakeholders to allow them to seek information about Open Data Triage processes as well as to provide them with the opportunity to challenge decisions and escalate issues.

3.111. The Licensees must keep under review its collection of available Data Assets, their associated Metadata and/or Software Scripts used to process Data Assets for risks or sensitivities and must mitigate these as they arise.

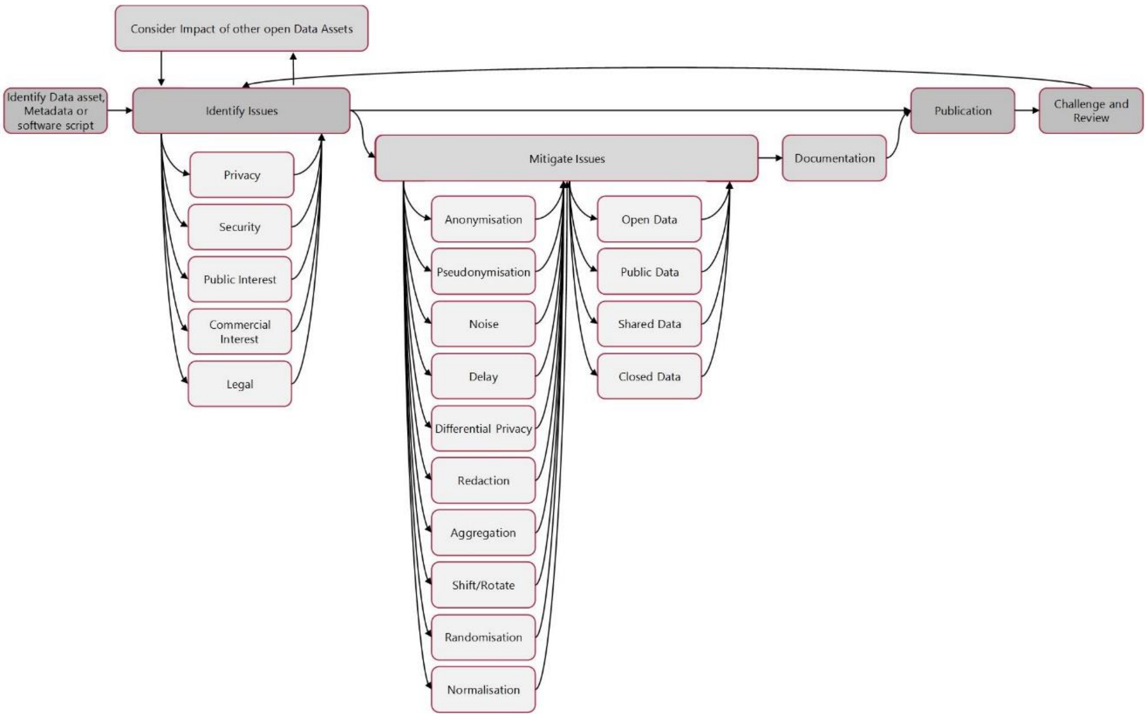
Techniques

3.112. When conducting Open Data Triage, it is more sustainable and flexible over time to focus on providing original data and the Software Scripts used to process it, rather than to focus on triage of data after it has been processed. This is because data has so many applications and so the Data Custodian will find itself having to manage a much more complex collection of Data Assets through this approach, further, the act of making Software Scripts open and available removes ambiguity about how data is treated and so will allow for Data Users to self-service many of their questions, improving efficiency.

3.113. The Open Data Triage process considers sensitivity themes such as personal privacy, security, legislation/regulation requirements, commercial needs that if not protected would have a negative impact on end-consumers and the Public Interest.

3.114. Legislation and regulation can also be enabling reasons that give Data Custodians the confidence that they have the legal right to process and publish data, for example, the Freedom of Information Act, Environmental Information Regulations, specific regulatory Licence Conditions (or Code obligations) all in their own context create lawful responsibilities for treating data as Open Data.

3.115. The diagram below is a high-level representation of the proposed process, more detail is provided about the steps in the following subsections.



Identification of Discrete Datasets

3.116. The goal of Open Data Triage is to identify where issues exist that would prevent the fully open publication of data in its most granular format and address these issues in a way that mitigates these issues, while maintaining as much utility (and therefore potential benefits) from the datasets as possible.

3.117. Effective Open Data Triage processes do not take a binary view of whether Data Assets can be open or not, but instead adopt the stance of "what are the least number of changes I need to to this Data Asset before I publish it" (and equivalently for Metadata and Software Scripts used to process Data Assets).

3.118. To make this process manageable for the Data Custodian, a good first step in this process is to identify thematic, usable datasets that can be joined if required rather than general data dumps.

3.119. In this context, a 'thematic, usable dataset' means a discrete collection of data that relates to a focused, coherent topic but provides enough information to be of practical use. Data Custodians should consider:

- 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, industry, etc.)

3.120. For example, an infrastructure construction company may have data about the construction and operation of various building projects across a number of countries. It is sensible to split the data into operational and construction data and then group by type of construction (public space, office building, residential building, etc.), geographic region and year of construction.

3.121. The approach described above minimises the risk that the size and complexity of datasets results in issues that are not correctly identified. It also reduces the risk that an issue in one part of the dataset results in the whole dataset being made less open or granular therefore maximising the amount of useful data that is openly available in its most granular form. For example, providing complete output from a data warehouse in one data dump could contain information about customers, employees, financial performance, company Key Performance Indicators (KPIs), etc. all of which would present issues that would mean the data needs to be modified or the openness reduced. Whereas extracting tables (or parts of tables) from the data warehouse would provide a more granular level of control which enables individual issues to be identified and addressed accordingly which would in turn maximise the data which is made openly available.

Identification of Issues

3.122. Once a thematic, usable dataset has been identified the data controller should assess the dataset to identify if there are any issues which would prevent the open publication of the data in its most granular format. Next, it is helpful to identify the potential issues which might limit the openness or granularity of the dataset.

3.123. In table 12 below, a range of issue categories are outlined that should be carefully considered. Some of these categories will directly relate to existing triage processes that already exist in organisations, but others may require the adaptation of existing processes or creation of new processes to provide a comprehensive solution.

Table 12: Issue categories relating to Open Data Triage

Issue Category	Description	Guidance
Peoples' rights to personal privacy ("Privacy")	Data that relates to a natural person who can be identified directly from the information in question or can be indirectly identified from the information in combination with other information.	This should be a familiar process as the General Data Protection Regulations (GDPR) introduced a range of requirements for organisations to identify personal data[95] and conduct Data Privacy Impact Assessments[96] (DPIA). The ICO has a wealth of advice and guidance on these topics, including definitions of personal data and DPIA templates. It is important that Open Data Triage is used to effectively identify privacy issues and ensure that any data which is released has been appropriately processed to remove private information and retain customer confidence in the product, service or system.
Security Needs ("Security")	Data that creates incremental or exacerbates existing security issues which cannot be mitigated via sensible security protocols such as personnel vetting, physical site security or robust cyber security.	Companies and organisations that own and operate infrastructure should already have a risk identification and mitigation program to support the protection of Critical National Infrastructure (CNI). The Centre for the Protection of National Infrastructure97 (CPNI) have advice and guidance for organisations involved in the operation and protection of CNI. Outside of CNI, organisations should assess the incremental security

[95] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data/
[96] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/#dpia5
[97] https://www.cpni.gov.uk/

		risks that could be created through the publication of data. Organisations should consider personnel, physical and cyber security when identifying issues and identify if the issue primarily impacts the publishing organisation or if it has wider impacts. Issue identification should take into account the existing security protocols that exist within an organisation and flag areas where the residual risk (after mitigation) is unacceptably high. Note, where that information contained within a Data Asset or software script is already publicly available via existing means (such as publicly available satellite imagery) the security issue assessment should consider the incremental risk of data publication using the existing situation as the baseline.
Negative impact on the Public Interest (“Public Interest”)	Data that is likely to drive actions, intentional or otherwise, which will negatively impact the consumer	Organisations should consider how a Data Asset could be used to drive outcomes to the Public Interest such as by enabling manipulation of markets, embedding bias into products or services incentivising actions that are detrimental to delivery of the Net Zero target. The Open Data Institute[98] (ODI) Data Ethics Canvas[99] may be of use when identifying potential negative impacts to the Public Interest.

[98] https://theodi.org/
[99] https://theodi.org/wp-content/uploads/2019/07/ODI-Data-Ethics-Canvas-2019-05.pdf

Commercial requirements that, if not protected, will have a negative impact on Products and Services for end-consumers ("Commercial Interest")	Data that typically 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 and where the publication of that data would effect the decisions of that business ultimately leading to a worsening of the Products and Services available to end-consumers.	Commercial data relating to the private administration of a business (HR, payroll, employee performance, etc.) may be subject to an Open Data Triage process, but is expected to not be made Open Data. However, de-sensitised versions of the data (eg aggregated statistics about pay awards between genders) might still be made available. Data which does not relate to the administration of the business but has been collected or generated through actions which are outside of the organisation's legislative or regulatory core obligations and funded through private investment may also have legitimate reason to be de-sensitised, shared with only a limited audience or even closed. Where an organisation is a regulated monopoly, special consideration should also be given to the privileged position of the organisation and the duty to enable and facilitate competition within their domain and along the Product and Service supply chains that they operate in. Where datasets contain Intellectual Property (IP) belonging to other organisations or where the data has been obtained with terms and conditions or a data licence which would restrict onward publishing this should also be identified. However, there is an expectation that organisations should be migrating away from restrictive licences / terms and conditions that restrict onward data publishing and sharing where possible.

Obligations from legislation and/or regulation (“Legal”)	Specific legislation or regulation exists which prohibits the publication of data.	Organisations should have legal and regulatory compliance processes which are able to identify and drive compliance with any obligations the company has. The types of legislation that may be relevant includes: Utilities Act 2000 Electricity Act 1989 Gas Act 1986 / 1995 Competition Act 1998 Enterprise Act 2002 Enterprise and Regulatory Reform Act 2013 Data Protection Act 2018 General Data Protection Regulation (GDPR) It is beneficial to the government and its developed bodies to be able to clearly understand which parts of the legal framework are preventing the treatment of data as Open Data so that the ongoing appropriateness (or not) of statutory instruments can be reviewed and revised more effectively.

Consider the impact of other Open Data Assets

3.124. When assessing the sensitivity of data, thought should be given to the other datasets which are already openly available and the issues that may arise from Data Users combining these datasets. Organisations should consider where there are datasets outside of their control which, if published, could create issues which would need to be mitigated. Special consideration should be given to datasets which share a common key or identifier, this includes but is not limited to:

- subject reference (e.g. Passport Number),
- technical reference (e.g. Serial Number),
- time (e.g. Universal Time Coordinated or UTC),
- space (e.g. Postcode or Property Identifier)

3.125. Over time markets will continually develop and public attitudes will change, there is therefore expected to be a need to revise Open Data Triage assessments over time. For example, a Data Asset that was initially deemed too sensitive to be released openly in its most granular form could be rendered less sensitive due to changes to market structure or change in regulatory obligations. Equally, a Data Asset that was published openly could become more sensitive due to the publication of a related Data Asset or technology development. Data Custodians will need to act on evolution in the market and will need to coordinate with other Data Custodians, who are also making data available.

Mitigation of Issues

3.126. Where an Open Data Triage process identifies a sensitivity issue, the aim is to mitigate the issue, such as by processing the data to modify it in a way that removes the sensitivity or to limit which stakeholders can have access to it, whilst still maximising its utility to stakeholders. To maximise its utility, the Data Custodian will usually have to work with stakeholders to understand their needs and so minimise the extent that it prevents these being met as the de-sensitisation takes place.

3.127. Open Data that has been processed to include some redaction may be preferable to only sharing data with a select few stakeholders. It may be required that both of these solutions are delivered. ie in some cases, the objectives of the prospective data users might create requirements which cannot be resolved by a single solution so it may be necessary to provide different variations or level of access, for example providing open access to a desensitised version of the data for general consumption alongside shared access to the unadulterated data to a subset of known Data Users.

Modification of Data

3.128. Modification of data can reduce its sensitivity, therefore enabling this de-sensitised version of the data to be treated as Open Data. There are a wide variety of possible ways to modify data that can be used to address different types of sensitivity.

Table 13: Ways to modify data for sensitivities

Technique	Description	Example Application	Commentary
Anonymisation	Removing or altering identifying features	**Privacy** An organisation has a licence condition to collect certain data about individual usage of national infrastructure. The data is collected about individual usage on a daily basis and could reveal information about individuals if it was to be released openly. By removing identifying features such as granular location and individual reference it could be possible to successfully anonymise the data such that individuals cannot be re-identified so the data could be made openly available.	Simple anonymisation can be very effective at protecting personal data but it needs to be undertaken with care to minimise the risk of re-identification. Anonymisation techniques can be combined with other mitigation techniques to minimise this risk. The UK ICO have provided an anonymisation code of practice which should be adhered to[100],[101]

[100] https://theodi.org/article/how-do-organisations-perceive-the-risks-of-re-identification/

[101] https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

Technique	Description	Example Application	Commentary
Pseudonymisation	Replacing identifying features with a unique identifier that retains the reference to an individual whilst breaking the link with the 'real world' identity.	**Privacy** An organisation (with permission) collects data about how customers use a web service. This data is used to diagnose problems where there are issues with the website operation. Replacing the customer name and address with a random unique identifier that allows the behaviour that led to an issue to be analysed whilst protecting the identify of an individual user.	Pseudonymisation is distinct from Anonmysation as it is possible to consistently identify individuals but not link this to a specific, named person. Pseudonymisation should be used carefully as it is often possible to utilise external datasets and data analysis to match identifiers and trends such that the individual can be re-identified. for example, it may be possible to analyse the website usage patterns (times, locations, device type, etc.) and cross reference with other personally identifiable datasets (social media posts, mobile positioning data, work schedules, etc.) to identify an individual with a sufficient level of confidence. Again the ODI and ICO provide useful guidance in this area.[102], [103]

[102] https://theodi.org/article/how-do-organisations-perceive-the-risks-of-re-identification/
[103] https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

Technique	**Description**	**Example Application**	**Commentary**
Noise	Combining the original dataset with meaningless data	**Commercial Interest** An organisation collects information about how individuals use a privately built product or service (e.g. a travel planner). This data could be of great use for the purposes of planning of adjacent system (e.g. energy system or road network) but releasing the anonymised, granular data would given competitors a commercial advantage. By introducing random noise into the dataset in a way that ensure that the data remains statistically representative but the detail of individuals is subtly altered the data can be made available whilst reducing the commercial risk.	Introducing noise to data in a way that successfully obfuscates sensitive information whilst retaining the statistical integrity of the dataset is a challenging task that requires specialist data and statistics skills. Consideration needs to be given to the required distribution, which features the noise will be applied to and the consistency of application.

Technique	Description	Example Application	Commentary
Delay	Deferring publication of data for a defined period	**Security** An organisation operates a network of technical assets some of which fail on occasion. If the data related to those assets was made available innovators could help to identify patterns which predict outages before the occur and improve the network stability. However, the data could also be used to target an attack on the network at a point which is already actively under strain and cause maximum impact. By introducing a sufficient delay between the data being generated and published the organisation can mitigate the risk of the data being used to attack the network whilst benefiting from innovation.	Delaying the release of data is a simple but effective method of enabling detailed information to be released whilst mitigating many types of negative impact. However, it may be necessary to combine this with other mitigation techniques to completely mitigate more complicated risks. Delay can work in many contexts, sometimes a delay of only a few seconds can remove the commercial sensitivity associated to data, in other cases a delay of many years might be appropriate to de-sensitise information about the business operations of a regulated monopoly, while still eventually enabling Public Interest benefits to be gained about topics such as market design and research into effective methods of conducting economic regulation.

Technique	Description	Example Application	Commentary
Differential Privacy	An algorithm or model which obscures the original data to limit re-identification	Privacy An organisation collects rich data from customers which is highly valuable but sensitive (e.g. email content). The sensitivity of this data is very high but the potential for learning is also very high. Differential privacy enables large amounts of data to be collected from many individuals whilst retaining privacy. Noise is added to individuals' data which is then ingested by a model. As large amounts of data are combined, the noise averages out and patterns can emerge. It is possible to design this process such that the results cannot be linked back to an individual user and privacy is preserved.	Differential privacy is an advanced technique but can be very effective. It is used by top technology firms to provide the benefits of machine learning but without the privacy impact that is usually required. Sharing a model can be a highly effective way of enabling parties to access the benefit of highly sensitive, granular data but without proving direct access to the raw information. However, this is an emerging area so carries some complexity and risk[104].

[104] https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Technique	Description	Example Application	Commentary
Redaction	Removing or overwriting selected features	**Security / Legislation and Regulation** An organisation maintains data about a large number of buildings across the country and their usage. Within the dataset there are a number of buildings that are identified as Critical National Infrastructure (CNI) sites which are at particular risk of targeted attack if they are known. In this case it is possible to simply redact the data for the CNI sites and release the rest of the dataset (assuming there is no other sensitivity). Note, this approach works here because the dataset is not complete and therefore it is not possible to draw a conclusion about a site which is missing from the data as it may simply have not been included.	Redaction is commonplace when publishing data as it is a very effective method of reducing risk. However, care needs to be taken to ensure that it is not possible to deduce something by the lack of data. In general, if the scope and completeness of data is sufficient that the lack of data is noteworthy then redaction may not be appropriate. e.g. An authoritative map which has a conspicuous blank area indicates the site is likely to be of some interest or importance.

Technique	Description	Example Application	Commentary
Aggregation	Combining data to reduce granularity of resolution, time, space or individuals	**Commercial Interest** An organisation collects information about the performance of their private assets which form part of a wider system (e.g. energy generation output). This data could be of great use to the other actors within the system but releasing the data in its raw format may breach commercial agreements or provide competitors with an unfair advantage. By aggregating the data (by technology, time, location or other dimension) the sensitivity can be reduced whilst maintaining some of the value of the data.	Aggregation is effective at reducing sensitivity but can significantly reduce the value of the data. It may be worth providing multiple aggregated views of the data to address the needs of a range of stakeholders. Where aggregation is the only effective mechanism to reduce sensitivity organisations may want to consider providing access to aggregated data openly alongside more granular data that can be shared with restricted conditions. Note, aggregating data which is of a low level of accuracy or quality can provide a misleading picture to a potential user. Custodians should consider this when thinking about the use of data aggregation and make potential users aware of potential quality issues.

Technique	Description	Example Application	Commentary
Shift / Rotate	Altering the position or orientation of spatial or time series data	**Privacy** An organisation collects data on how their customers use a mobile product including when and where. This movement data could be of value to other organisations in order to plan infrastructure investment but the data reveals the patterns of individuals which cannot be openly published. The initial step is to remove any identifying features (e.g. device IDs) and break the movement data into small blocks. Each block of movement can then be shifted in time and space such that they cannot be reassembled it identify the movement patterns of individuals. This means that realistic, granular data can be shared but the privacy of individuals can be protected.	Shifting or rotating data can be useful to desensitise spatial or temporal data. However, it is important to recognise context to ensure that the data makes sense and cannot be easily reconstructed. For example, car journey data will almost always take place on roads and therefore rotation can make the data nonsensical and it can be pattern matched to the underlying road network with relative ease.

Technique	Description	Example Application	Commentary
Randomisation	Making arbitrary changes to the data	**Security** An organisation may be the custodian of infrastructure data relating to a number of sensitive locations such as police stations or Ministry of Defence (MoD) buildings. The data itself is of use for a range of purposes but making it openly available could result in security impacts. Randomising the data (generating arbitrary values) relating to the sensitive locations (rather than redaction) could reduce the sensitivity such that it can be open.	Randomisation can be very effective to reduce sensitivity but it is also destructive so impacts the quality of the underlying data. Randomisation of data differs to adding Noise to data in that randomisation replaces the original data, whereas Noise is included in addition to the original data.
Normalisation	Modifying data to reduce the difference between individual subjects	**Negative Consumer Impact** An organisation provides and collects data about usage of a product in order to diagnose problems and optimise performance. The data has wider use beyond the core purpose, but the associated demographic data could result in bias towards certain groups. By normalising the data (reducing variance and the ability to discriminate between points) it is possible to reduce the ability for certain factors to be used to differentiate between subjects and hence reduce types of bias.	Normalisation is a statistical technique that requires specialist skills to apply correctly. It may not be enough on its own to address all sensitivities so a multifaceted approach may be required.

3.129. Data Custodians should consider the impact of data modification on the usefulness of the data and seek to use techniques which retain the greatest fidelity of data whilst mitigating the identified issue. Given the diversity of data and variety of use cases it is not possible to define a definitive hierarchy of preferred data modification techniques. Consideration should be given to making data available using more than one method or preference given to techniques that have limited impact on the overall scope and accuracy of the data over those which make substantial, global changes. e.g. Redaction of sensitive fields for <1% of records is likely to be better than dataset wide aggregation.

3.130. The publication of Open Data Triage processes provides transparency and promotes consistency across the sector and enables challenge. Data custodians should ensure that when releasing Software Scripts used to process Data Assets that this does not enable users to reverse engineering the raw data or sensitive information. For example, if noise is being added then the noise pattern should not be reproducible or possibly redacted from the publication.

Level of Access

3.131. While identifying mitigation techniques, the Data Custodian should also consider how open the resulting data can be. Where the mitigation has been successful the data can be treated as Open Data and so published for all stakeholders to use freely. However, if the nature of the data means that it is only valuable in its most granular form it may be necessary to reduce openness but make more sensitive versions open to certain stakeholder groups.

Table 14: Levels of Access description

Level of Access	Description
Open	Available for all stakeholders to use, modify and distribute with no restrictions
Public	Publicly available to all stakeholders but with some restrictions on usage
Shared	Available to a limited group of stakeholders possibly with some restrictions on usage
Closed	Data is only available within the Data Custodian's organisation, and it may be limited even within that.

The above table is based on the ODI data spectrum[105].

Balancing Openness, Modification and User Needs

3.132. A key factor to consider is the needs of the potential data users. Initially, there may be value in providing aggregated summaries of data which can be made entirely open but as new user cases and user needs emerge, we may find that access to more granular data is required which necessitates a more sophisticated mitigation technique or a more granular version of the data which is shared less openly. In some cases, it may be prudent to make multiple versions of a Data Assets, their associated Metadata and Software Scripts used to process Data Assets available to serve the needs of a range of Data Users.

Complex Issues

3.133. The goal of a Presumed Open approach is to make data as accessible as possible, this enables innovators to opportunistically explore data and identify applications that are not obvious. Wherever possible, the Data Custodian should seek to identify a mitigation approach which addresses all issues whilst maximising access to the most granular data.

3.134. In some cases, the number and diversity of issues may be so great that the mitigation or reduction of openness required to address all of the issues simultaneously is deemed too detrimental to the overall value of the data. In these cases, the Data Custodian could consider the user needs and individual use cases to help guide the mitigation strategies. For example, it may be possible to provide aggregated data openly for the purposes of statistical reporting but more granular data to a set of known participants via a secure data environment for another use case.

[105] https://theodi.org/about-the-odi/the-data-spectrum/

Documentation

3.135. When documenting issues identified by an Open Data Triage process, a description of the issues should be given as well as the mitigation option(s) available, a guide for this is as follows:

Table 15: Documenting issues during Open Data Triage

Issue Category	Description of Issue
Privacy	A description of the data and identification of features that contain personal data with additional flag for sensitive personal data.
Security	Where possible, a description of the data and features which cannot be published. There may be some cases where acknowledging the data exists may represent a security risk, in which case the relevant authority (eg regulator or government department) should be consulted.
Public Interest	A description of the data, sensitive fields or processing methodology and overview of the likely negative impact. Whilst it might not be possible to describe the likely negative impact in detail there should be some indication of the cause of the sensitivity e.g. market manipulation.
Commercial Interest	A description of the data along with an outline of how this impacts commercial interest, e.g. business administration data or justifiable private investment.
Legal	A description to the data and reference to specific clauses and articles of legislation or regulation that prohibits publication.

Examples

Identification of Discrete Datasets

3.136. The dataset that has been identified is electricity network substation connection capacity. This represents the likely connection headroom at every substation within an electricity distribution network's area. Monitoring data (substation and smart meters) is used to measure the current demand and network data is used to provide the maximum capacity.

3.137. Note - this is a scenario for the purposes of providing an example and does not represent an operational network monitoring solution.

Table 16: Exampe - Identification of Issues

Issue	Identification of Issues	Mitigation of Issues
Privacy	Substation connection capacity is not personal data. However, if the dataset includes total capacity and used or available capacity then in the rare case where a substation serves a single customer this could be an individual's private data. Amber	Analysis will take place before the data is released to identify any substations which serve a single premises. If these cases materialise the sensitive fields will be redacted and the data will be displayed as available capacity only to avoid personal data issues. Green
Security	If the data represents the live status of the network then a bad actor could use this information to inform an attack. Amber	Data can be delayed by 24 hours such that it is not possible to determine live status of the network. Green
Negative Consumer Impact	If an actor used the data to opportunistically request connections that utilise all of the available capacity it could drive up costs for future users, including for new housing, a cost which could be passed on to the consumer. Amber	Ensure there is a process in place to stop actors stockpiling capacity and distribute costs fairly. Green
Commercial Interest	None - the network is a monopoly player with the duty to deliver an efficient, competitive system. Green	Green
Legislation and Regulation	GDPR (if personal data is involved) Amber	See privacy mitigation. Green

3.138. Due to the availability of options to mitigate all of the issues above, the Data Asset can be processed to de-sensitise it and that de-sensitised version can be treated as Open Data and published.

3.139. The associated Metadata description can also be treated as Open Data and this will include details about which software script were used to de-sensitise the Data Asset. The software script used to process Data Assets is also published, as sharing its methodology does not create any sensitivities. In a more complex example, a longer data pipeline may need to be made available and this may require some additional documentation.

3.140. The Data Asset might be published using csv file format and the Data Assets, their associated Metadata and Software Scripts used to process Data Assets then hosted on an open data access platform with the Metadata format compatible with its being automatically registered with open data catalogues.

Challenge and Review

3.141. A part of the purpose of publishing information associated with Open Data Triage decisions is to ensure that stakeholders are enabled to challenge decisions where they do not agree with the rationale and evidence provided by the Data Custodian. Stakeholder challenges may not be to overturn a decision as a whole, but instead may be new ideas for how to approach the mitigation of sensitivities in ways that better protect the utility of Data Assets.

3.142. For example: An innovator indicates that for their specific use case they need to be able to access the data in near real time, but that a mitigating action to a sensitivity in a Data Asset is to delay its publication by a week. The innovator proposes an alternative mitigation, where there is no delay to publication, but where less fields of data are published in real-time. The Data Custodian adopts this new solution to de-sensitising the data and now both publishes a limited dataset in real-time and a more detailed dataset with a one week delay.

Templates

3.143. The following canvas (in pdf and pptx format) has been designed for use in a collaborative Open Data Triage workshop. They are designed to help structure a group discussion rather than be completed in isolation.

[Open Data Triage Canvas.pptx]
[Open Data Triage Canvas.pdf]

Documentation

3.144. The following documentation templates are provided to show what information should be recorded following the open data triage process. They are not final or authoritative and other systems or tools may be better suited to the needs of individual organisations.

[Open Data Triage documentation.xlsx]
[Open Data Triage documentation.csv]