

## **Underlying data publication: interim guidance**

Version:	0.4
Date created:	15/07/2010
Policy official:	David Pullinger
Date last updated:	19/07/2010
Date issued:	Tbc
Lead official:	David Pullinger
Guidance number:	TG135

### **Purpose**

**This guidance is for all public sector communicators, website managers and policy teams. It covers the minimum requirements for releasing data published in documents in a re-usable format as well as the optimum form.**

## Table of contents

<b>Introduction .....</b>	<b>2</b>
Aims and objectives .....	2
Audience .....	2
Scope .....	2
Background .....	2
<b>Requirement .....</b>	<b>4</b>
<b>Identifying underlying data .....</b>	<b>5</b>
Types of underlying data .....	5
Finest level of granularity possible .....	5
Collecting data together .....	5
<b>Releasing data.....</b>	<b>6</b>
A stable and permanent Internet location for the data release.....	6
A good user experience .....	6
Timely data release.....	6
Findable data .....	6
Copyright and re-use.....	6
Registering the release in data.gov.uk.....	7
<b>Provenance and information about the data .....</b>	<b>8</b>
Give information about the data .....	8
A minimum set of fields about the data .....	8
Additional fields that are useful .....	8
<b>Data formats .....</b>	<b>9</b>
Enable machine-processable data .....	9
Use open standardized formats with easily extractable data .....	9
Mandated formats for specific data .....	10
Minimal format for data release.....	11
Optimum for data release.....	11
<b>Oversight and support for data release .....</b>	<b>13</b>

## Introduction

### Aims and objectives

The Government aims to release all underlying data in a re-usable way from its published information. Specifically, 'From July 2010, government departments and agencies should ensure that any information published includes the underlying data in an open standardized format.' (PM's letter of 29 May 2010 to Secretaries of State)

It would be possible for each public body to release the data in such a way that it is collectively no more useful than in the publications themselves. The objective of this guidance is to identify data that should be released and to ensure that the data is easy to work with using common tools.

This guidance is interim guidance pending further consultation.

### Audience

This guidance is for public sector communicators, website managers, and policy teams. It applies to all central government departments both Ministerial and non-Ministerial, their agencies and non-departmental public bodies (NDPBs) and arms-length bodies. It should be considered as good practice for the rest of the public sector, including by local government and major public services.

### Scope

#### In scope:

- Underlying data from documents published by the public sector
- Minimal requirements for releasing underlying data
- Data which must be released in pre-specified formats
- Format choices

#### Out of scope:

- Real-time data
- Publication of large stand-alone datasets
- Detailed technical information
- Legal aspects including, anonymisation and data protection

### Background

Government publishes many official documents containing data, including research and statistics, annual reports, policy documents and accountability statements. The data contained in these reports is useful and interesting to other parties, although not easily extracted, particularly by computer processing. Underlying data should therefore be published online so that they are extractable and processable by machines without the need for direct human intervention.

This guidance identifies the kind of data which should be considered for data release in a reusable form. It also describes the formats by which underlying data should be released to be of most usefulness to third parties, allowing the data to be processable and so used to create new analyses, services, communications and commercial products and businesses.

There is a further expectation, but not mandate, that public bodies will enable the data to be not only processable by machines but also understandable and interpretable by people. This may take the form of visualizations or interactive tables. People should therefore be able to read and have a good user experience in understanding the data, either through summaries written in text with tables and diagrams or interaction and visualization tools.

## Requirement

1. Government departments, their agencies and arms-length bodies should:
  - release all the data associated with any publication;
  - as timely as is feasible;
  - in the finest level of granularity possible;
  - beginning no later than July 2010;
  - in an open standardized format;
  - that facilitates re-use both for intended and unintended purposes;
  - freely available on the Internet;
  - in a stable and permanent location;
  - with useful information about the release and the data;
  - with clear indication of permissible use (copyright statement);
  - and registered on [data.gov.uk](http://data.gov.uk).

## Identifying underlying data

2. Published information has underlying data can take many forms and is of wide-ranging interest and use to others. The types of data range from that held in tables through geographical information, regular reports structured in a common way (for example inspection reports) to policy statements concerning, for example, breakpoints in eligibility for benefits or tax bands.

### Types of underlying data

3. Rather than precisely identify the types of data that should be published, public bodies should follow the presumption of openness.
4. Publishers of official documents should therefore review all published information with the aim of identifying the different types of data and reviewing the appropriate way to release data.
5. In case of doubt, they should contact [publicdata@nationalarchives.gsi.gov.uk](mailto:publicdata@nationalarchives.gsi.gov.uk).

### Finest level of granularity possible

6. Underlying data is data used as a basis for whatever is reported in the publication. It should be released at the finest level of granularity possible, within legal constraints of non-disclosure of personal data.
7. Data should be available to support both the intended and unintended uses of the data. Redacted and processed data designed purely for the publication is not therefore acceptable.
8. It is the responsibility of the publisher to describe the state of the data, including the level of granularity at which it has been published.

### Collecting data together

9. Data pertaining to one theme should be released as a single dataset, even though there may be multiple tables or figures pertaining to it in a report or publication. In a typical report either all the data pertains to one dataset or they tend to break naturally into themes and subjects.

## Releasing data

### A stable and permanent Internet location for the data release

10. It is important that it is always possible to link back to the primary unmodified data source as provided by the government. Each dataset from an underlying publication should be located on its own Web page. This will allow the URL to be referenced and, through the UK Government Website Archive, permanently retrievable. As part of the terms and conditions of re-use, third parties are required to reference their source. (Technically this means that the URL location can act as a URI identifier a concept also called a globally unique identifier, GUID.)

### A good user experience

11. Either in the text of the publication or in the release of the underlying data, a human reader should be able to view and understand the data that is being released.

### Timely data release

12. The desire to release data in an optimal format should not delay publication. If necessary an interim format may be used, for example CSV, while the data is being further formatted.

### Findable data

13. To ensure findability, the pages on which the data may be found should be included in the website's XML Sitemap. This allows Web search engines to index the content and so people to find through search.
14. In addition, the robots.txt file should be checked to see that it is not inadvertently blocking access to the data.

### Copyright and re-use

15. A clear statement should be made as to the licence under which people can re-use data. This is done through the Crown Copyright terms and conditions, with allow people to take and re-use data provided they cite the source and do not misuse to deliberately mislead etc. The details are published in TG 113 Legal Issues <http://www.coi.gov.uk/guidance.php?page=164>

## **Registering the release in data.gov.uk**

16. The data release is considered published and so public if and only if registered on data.gov.uk website.
17. There will be a set of accounts that are authorized to register data on data.gov.uk. While automatic tools for this are being developed, please contact [publicdata@nationalarchives.gsi.gov.uk](mailto:publicdata@nationalarchives.gsi.gov.uk).

## Provenance and information about the data

### Give information about the data

18. All underlying data from a publication, whether intended for printing (for example in PDF format) or viewing online (for example HTML), should contain both information about the data itself ('metadata') and provenance information ('contextual metadata').

### A minimum set of fields about the data

19. The standard, *Meta-data standard for data.gov.uk registration*, is currently being developed and the core of the new standard will be the following:

Element name	Type
Identifier	Key
Title	String
Abstract	String
Department	Controlled vocab.
Contact	String
Contact e-mail	E-mail address
Licence	Controlled vocab.
Resource format	Extensible controlled vocab.
Resource URL	URL
Resource ID	ID
Publisher	Controlled vocab.

### Additional fields that are useful

20. In addition, some contextual information is very useful:
- The name of the publication in which the data was released
  - The date of the publication in which the data was released
  - A description of the data in non-jargon terminology
  - A description of the formats in which the data is released
  - Frequency of update (if relevant)
  - Geographical coverage

## Data formats

### Enable machine-processable data

21. In the context of data release, machine-processable means making any underlying data used in publications accessible for use by a computer-based process, not requiring human interpretation. At one level, all information and data available on computer-based devices are machine-processable. Word document files are machine-processable in the sense that Word and other compatible programs are able to interpret the data and present it as text on a screen. Similarly HTML is a standard that indicates how to display Web pages in a browser. The key aspect for release of *underlying* data is that it can be extracted from any particular format and re-used and re-purposed.
22. Data should be released in a way that is useful to others. There is always a trade-off in determining what aspects of information are made machine-processable. For example, minutes of meetings might or might not contain the geographical location of the meeting place in latitude and longitude, However the added cost of adding such coordinates to the data may not be justified. Some formats may be post-processed to add additional information and these are therefore to be preferred. The principle is to release as much as is possible.

### Use open standardized formats with easily extractable data

23. The choice of format is important in releasing public sector data as some formats are susceptible to becoming obsolete, have unintended technological limitations or retain restrictive licensing arrangements. The intention is that formats used do not rely on a single product or company in order to be used, are fully documented, are freely available to the public without any licensing, patent or other restrictions and also maximize potential uses.
24. Data should be released using open standardized formats for which the data is easily extractable.
  - By open standards is meant non-proprietary, being freely available and where public documentation of the standard is published.
  - By standard is meant those governed by international standards bodies, or adopted as industry standards.
  - By easily extractable is meant that there are DIY pre-existing tools and support for extracting the data.
25. The appropriate open standardized formats vary according to the kind of data being released. They will be expected to be one of the following:
  - **CSV:** A comma-separated values file is a simple text format for a [database](#) table. Each record in the table is one line of the text file. Each field value of a record is separated from the next with a comma. The first

line of the text file may be a header line containing a usable 'name' for each field.

- **RDFa** or Resource Description Framework – in – attributes is a World Wide Web Consortium (W3C) Recommendation that allows a Web page to have added metadata so that data can be extracted as RDF model triples by compliant user agents. This means that the same page of data can be both understood by human readers and by machines.
- **XML (Extensible Markup Language)** is a set of rules for encoding documents electronically. There are many specific uses of XML for data, for example for tables and mathematics.
- **RDF/XML** : The **Resource Description Framework** is a family of World Wide Web Consortium (W3C) specifications that has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax formats. The format used following the modelling is XML.

26. The organization releasing the data should not randomly choose one of these as they have different uses and benefits.

27. Such guidelines should be tempered with the principle of usability and usefulness. They should contain all the data and associated information required to make them both usable and useful. For example:

- CSV files may not be useful without a description of all the columns and rows making up the data table. This should be provided as a separate descriptive document or a data dictionary if necessary; the file itself should only contain one headline containing a usable 'name' for each field.
- Although PDF could be considered non-proprietary in being an open international ISO standard, PDF is not a good choice, even with embedded XML metadata (see below), as there remains a problem of locating and readily extracting the data.
- Although HTML is accepted as an international industry standard, it requires individual programming to interpret the HTML formats in order to get at data ('screen-scraping') and so could be considered acceptable only in the circumstance that there is a commitment never to change the Web pages.
- There are other formats, such as ODIF (Open Document interchange format) which is a free and open international standard document file format maintained by the ITU-T to replace all proprietary document file formats. This might or might not be easy to import depending on the exact way it is set up.

### **Mandated formats for specific data**

28. The standard proposes two situations for publishing data;

- where there is accepted and published government practice for a particular kind of data, then this should be used;
  - where there is no accepted nor published government practice for a particular kind of data, then one of the recommended formats should be applied.
29. If there are published standards for the particular kinds of data, then data should be released using these. The following is a list, to which new standards are being added:
- **Organizational data**, including lines of accountability.
  - **Statistical data**
  - **Financial data**
  - **Job vacancy descriptions**: All public sector job vacancy descriptions are to be released using RDFa, so that they can be re-used and repurposed by third parties. The standard is set in TG124 *Structuring information on the Web for re-usability*  
<http://www.coi.gov.uk/guidance.php?page=312>
  - **Consultation descriptions**: All government consultations complying with the Consultation Code should have descriptions in RDFa, so that they can be re-used and repurposed by third parties. The standard is set in TG124 *Structuring information on the Web for re-usability*  
<http://www.coi.gov.uk/guidance.php?page=312>
  - **Geographic data**
30. Although some of the types of data listed in the preceding paragraph are generally published as units independent of any document, if a document were to contain such information, for example statistical and financial, then the underlying data should be released using the mandated formats.

### Minimal format for data release

31. The minimal general format for any data is a CSV file with description of the data (descriptive column and row headings or data dictionary) along with an HTML summary describing the content.

### Optimum for data release

32. There is, for the reasons given above, no single optimum format for machine-processable, usable and useful data. However, the use of RDF offers three main advantages:
- If data is released in RDF, then it can quickly be moved to any other format including XML, JSON and CSV, using common tools, whereas this is not true for other formats.
  - There are users that work with and so prefer different formats, so the use of one which can readily be moved into other formats is to be preferred.

- If data is released in RDF, it can be post-aligned to other data (other vocabularies) broadening its use and allow aggregation and processing of data into new information and services.
33. The optimum is when an interface is provided that makes access to and use of the data very easy to other machines. This is done through the provision of an application programming interface (API), which is an interface implemented by a software program which enables it to interact with other software.
  34. APIs using Representational State Transfer (REST), a style of software architecture for distributed hypermedia, are to be preferred because standalone RESTful GET commands tend to be intrinsic to the Web, simple, and retrieve data and do not change it. The alternative use of SOAP, if not correctly implemented, can introduce potential problems as it encourages each application designer to define a new command vocabulary over the top of existing Web standard capability, thus making it harder for software to get easy access to the data.
  35. Further information will be added onto Civil Pages, Making Public Data Public community, and readers who are interested in some of the more technical aspects are encouraged to read Sir Tim Berners-Lee's paper: *Putting Government data online*:  
<http://www.w3.org/DesignIssues/GovData.html>

## Oversight and support for data release

36. The oversight committee for data release is the Public Sector Transparency Board, chaired by the Minister for the Cabinet Office.
37. The overall SRO for the government is The Director of Digital Engagement, Government Communications, Cabinet Office.
38. Each Department has appointed an appropriate SRO for their sector.
39. Each public body has a Senior Information Risk Officer (SIRO) to clear data for release should there be any concern over the nature of the data.
40. There is a cross-government Making Public Data Public Practitioner Group, which meets monthly. This is a representative group for each Ministerial and non-Ministerial government department representing their sectors. This is organized by The National Archives.
41. New processes, tools and information will be communicated in the open community on Civil Pages, Making Public Data Public.
42. Questions that cannot be answered by the available resources or recourse to the Practitioner in the appropriate department should be addressed to [publicdata@nationalarchives.gsi.gov.uk](mailto:publicdata@nationalarchives.gsi.gov.uk).