Designing URI Sets for the UK Public Sector

A report from the Public Sector Information Domain of the CTO Council's cross-Government Enterprise Architecture

Interim paper

Version 1.0

October 2009

Change Control

Version	Date	Editor	Note
1.0	09/10/2009	Paul Davidson, CIO, Sedgemoor District Council	This document is the result of a series of workshops organised by the Chief Technology Officer (CTO) Council's Information Domain during July and August 2009, and wider feedback.

Contents

Introduction

1. This paper is a part of the family of policy and guidance documents associated with the cross-Government Enterprise Architecture (xGEA). Further guidance will be created to cover related topics such as:
   - Defining concepts to share meaning
   - Publishing linked open data
   - Querying linked open data

2. This document defines the design considerations and guidance by which UK public sector Universal Resource Identifier (URI) sets should be developed and maintained. They are designed both to encourage those that definitively own reference data to make it available for re-use, and to give those that have data that could be linked, the confidence to re-use a URI set that is not under their direct control.

3. The document will be of direct interest to:
   - Owners of reference data in the UK public sector
   - Data owners who wish to improve the re-use of their data by incorporating URIs that they do not control
   - Solution providers to the UK public sector

4. Some definitions and frameworks are laid out to define the types of resources that URIs can name, and the relationships between those types. A number of principles are then proposed against which a series of design considerations are made.

   These include:
   - Choosing the right domain for URI sets
   - The path structure for URIs
   - Coping with change and the passage of time
   - How to 'look up' a URI
   - The quality characteristics that apply to all URIs within a set
   - Machine-readable and human-readable formats
   - The governance arrangements necessary to allow the confidence to use and re-use UK public sector URIs

How URIs can be used to publish public sector reference data

5. Typically, government departments and agencies keep a list for each type of 'Thing' that they are responsible for, or handle in some way, and associate an identifying reference to each entry on the list. They then make use of that identifier as they make statements about the 'Thing' in their data. The lists therefore contain the 'Reference Data' that provide a common meaning and common identifier to refer to the same 'Thing' within that department or agency.

6. URI sets provide an opportunity to share common meaning and common identifiers across the public sector, and with the public, to join-up otherwise disparate data from many sources. Those that have confidence that the set is fit for their purpose are then likely to re-use it, rather than create their own.

7. Universal Resource Identifiers (URIs), a component of the World Wide Web, provide one means of uniquely naming a 'Thing' (or 'Resource'). The principles of 'Linked Open Data' rely on the RDF data model, where statements are made about resources, identified by URI(s).

8. URI sets will be an integral component of a UK Public Sector Information Architecture that supports many goals including the release of government data, reduced duplication, and increased information sharing towards transforming government services.

9. URI sets can be published by the UK public sector to provide comprehensive and reliable identifiers for 'Things' such as schools, roads, legislation, locations, projects, events and so on. Where the quality of these sets can be described consistently, other data owners will have the confidence to re-use them in their own data, leading to a web of data that can be linked, queried, and aggregated.

10. The existing UK public sector standards for metadata and 'findability' work well when applied to documents, but are not sufficient to support a 'Web of Data', where each individual statement can be queried and linked.

The need for design rules and guidance

11. As at September 2009, there are only a handful of early adopters of URI sets in the UK public sector, such as
    - BBC
    - Ordnance Survey
    - Office of Public Sector Information

    It is noticeable that, while each has faced similar design issues and choices, the implementations are quite different.

12. Much of the design in this paper is based on established and emerging good practice, whereas some implementation decisions are made to meet the specific needs of the UK public sector. In brief, these include:

    - Use of **data.gov.uk** as the domain to root those URI sets that are promoted for re-use
    - Organisation of URI sets into 'sectors' (e.g. education, transport, health) with a lead department or agency
    - Consistent use of metadata to describe the quality characteristics of each URI set

Evolution of the design considerations

13. This document is the result of a series of workshops organised by the Chief Technology Officer (CTO) Council's Information Domain during July and August 2009, and wider feedback to early drafts.

14. As URI sets are built and trialled, some scenarios may emerge that suggest that a rule may not be appropriate in some circumstances. Similarly, it may become apparent that there is value in considering that a piece of guidance should become a rule. It can therefore be expected that the interim design will be tested, challenged and proved by some early adopters, leading to a refresh.

15. Some scenarios, such as defining locations, may not fit well with the general principles of naming resources in this way. A further refresh of the design will illustrate how various scenarios are incorporated and aligned with sector-specific approaches to publishing reference data.

16. A web-based community of practice will be used to provide supporting:
    - Technical implementation guidance and bindings
    - Worked examples
    - Glossary of terms and definitions
    - Links to the published material that was used to support this guidance
    - Links to further material and good practice

Definitions, frameworks and principles

Types of URI

1. URIs can be used to name:

Type of Resource	Type of URI to name the resource	Definition / Scope
Real-world 'Things'	Identifier URI	These are the physical and abstract 'Things' that may be referred to in statements. Example of physical real-world 'Things' A school, a person, a road Example of abstract 'Things' A government sector, an ethnic group, an event Documents or 'works' are also examples of real-world things that can be named in this way as distinct from the content that they contain. Real-world Things can be referred to as 'Things' (with a

		capital 'T') A real-world 'Thing' cannot be found on the web, whereas information about it can. It is important, therefore, to be able to distinguish between a real-world 'Thing', as distinct from information about it, when making statements that refer to it.
Information on the web about real-world 'Things'	Document URI	These name the documents that are located on the web which are explicitly linked by the publisher of each 'Identifier URI' to provide information about real-world things.
	Representation URI	Where the publisher of a 'Document URI' provides more than one format, each format may be separately named by a Representation URI. Depending on the formats, some Representation URIs may name documents which are machine-readable and can therefore provide further links about the named resource.
Index of each of the identifiers within a set	List URI	These provide a list of the Identifier URIs that are contained within a set.
Definitions of concepts	Ontology URI	Whereas a real-world 'Thing' identifies an individual instance of that thing, there is also a need to provide a definition of the concept. The 'Ontology URI' could be looked up to give that definition.
Relationships between things	Ontology URI	Each part of an RDF statement can be named using a URI. This includes the relationship between real-world 'Things'. The 'Ontology URI' will then give a link into an ontology that can provide further reasoning about the relationship and the concepts that can be related using it.

What is a URI set?

2. For the purposes of this document, the term 'URI set' is to mean a collection of reference data published using URIs, about a single concept, governed from a single source. For example, each of schools, roads, legislation, would be a separate URI set.

3. An additional type of URI is required to name the URI set itself:

URI Set	Set URI	A type of Identifier URI that names the URI set and can be resolved to provide the quality characteristics of the set.

Design principles for public sector URI sets

4. The following principles have been derived from existing good practice and revised to meet the challenges for UK public sector URI sets:

Principle	
Use HTTP so that URIs can be resolved	MUST
Use a consistent path structure to explicitly indicate the type of URI	RECOMMEND
The publisher will make it clear whether the set is promoted for re-use by other parts of government and/or the public	MUST
Public sector URI sets should publish their expected longevity, and potential for re-use	MUST
Those public sector URI sets that are promoted for re-use should be designed to last for at least 10 years	RECOMMEND

Where more than one Representation URI is available, provide a Document URI where Content Negotiation can be used to provide the most appropriate representation	RECOMMEND
Avoid exposing the technical implementation of a URI in its structure	RECOMMEND
As a minimum, provide a machine-readable Representation URI	MUST
If appropriate, provide a human-readable Representation URI in HTML	RECOMMEND
Provide a means of discovering each of the available Representation URIs for a single Document URI	RECOMMEND
A URI set will publish its authorisation, authentication, and data quality characteristics using a common vocabulary	MUST
A URI structure will not contain anything that could change, such as session IDs	MUST
A URI path structure will be readable so that a human has a reasonable understanding of its contents	RECOMMEND

5. When considering which part of the UK public sector should set up URI sets:

Principle	Considerations
The department or agency responsible for a real-world 'Thing' should also be responsible for defining it and naming instances of it, on behalf of the appropriate sector	URIs should be organised into sectors with a lead department or agency. Lead departments/agencies should engage with stakeholders to ensure that the set is of sufficient quality to meet a wide range of purposes.
URIs from a set that is promoted for re-use should not contain the name of the department or agency currently responsible for it.	This copes with machinery of government changes where a department or agency may cease or have its scope changed.

Designing a URI set

What will a URI set contain?

1. A URI set will contain:
   - A URI to name the set and describe its quality characteristics
   - Each of the Identifier URIs for the real-world 'Things' in a single concept
   - Optionally, Ontology URI to define the scheme's concept and relationships
   - Optionally, List URI to list the Identifier URIs contained in the set

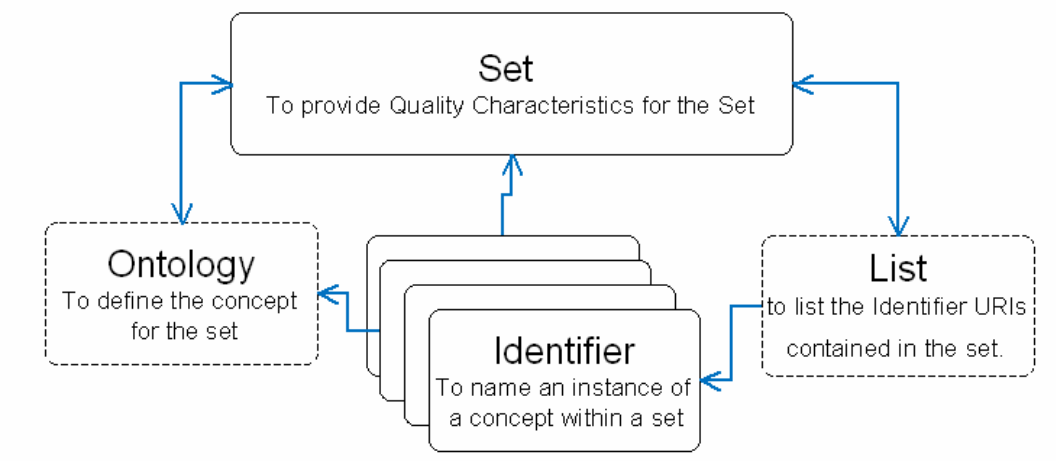


Figure 1: URIs that make up a set

2. Each type of URI will provide a means of looking up data on the web about the resource that the URI names, using:

- a Document URI
- at least one machine-readable Representation URI for each Document URI

3. The publisher of a URI set will implement facilities to provide mechanisms to:
   - Lookup an Identifier URI and be redirected to its Document URI
   - Discover each of the Representation URIs that are available for a Document URI
   - Get the most appropriate Representation URI for a specified application

4. Further guidance about how data can be looked up via Document and Representation URIs is provided later in this document.

Quality characteristics to publish for a URI set

5. The following information should be provided as a part of the metadata to describe the URI set:

Concept definition	Either by: • an Ontology URI that resolves to a machine-readable definition • as human-readable metadata
Relationships to other URI sets	To highlight other associated URI sets
Provenance	To describe the source and purpose of the reference data
Official status	To describe the range of statuses of the identifiers that that are contained in the set
Accuracy	To describe the closeness to the truth that the set attempts to achieve
Completeness	To describe the degree to which the Identifier URIs are a complete set against the definition of the concept
Timeliness	To describe the time-lag between a change to a real-world 'Thing' being applied to the URIs in the scheme
License Terms	To describe the terms of use for the URI set
Intended longevity	To provide a guarantee of persistence of the set
Intended audience	To describe who may confidently use the set. This provides a means of marking the set as being promoted for re-use
Representations available	To describe the range of file formats of Representation URIs in the set

6. Guidance as to the bindings and vocabularies used to express this metadata will be provided via the associated community of practice. This is likely to take the form of an RDF binding to an application profile of the e-Government Metadata Standard (eGMS) and/or Dublin Core properties and classes, or to sector specific metadata profiles, for example INSPIRE as applied to location.

How should URI schemes be discovered?

7. A mechanism will be required to provide a single query-point to discover each UK public sector URI set that is promoted for re-use. Guidance as to the method of discovery will be provided via the associated community of practice. This is likely to use established methods such as voiD[1].

8. Publishers may wish to offer a facility to subscribe to their set so that they can alert consumers to changes, improvements, etc.

Choosing the right domain for URIs

General

1. When considering the domain to root a URI set in:
   - the publisher will require content control of the sub-domain that it ultimately resolves to
   - the domain will have appropriate service-levels and scalability for resilience and performance

[1] voiD – vocabulary of interlinked Datasets, see - http://rdfs.org/ns/void/

Requirements for URI sets that are promoted for re-use

2. In addition, where a URI set is promoted for re-use, the following considerations apply to find a balance for central and federated components:
   - Flexibility and readability
   - Administrative burden
   - Infrastructure costs

3. In particular, the domain will:
   - Expect to be maintained in perpetuity
   - Not contain the name of the department or agency currently defining and naming a concept, as that may be re-assigned
   - Support a direct response, or redirect to department/agency servers
   - Ensure that concepts do not collide
   - Require the minimum of central administration and infrastructure costs
   - Be scalable for throughput, performance, resilience

4. The choice of domain should provide the confidence to the consumer, that the URI set has met minimum quality criteria, including implementing these design considerations. In other words, the domain itself should convey an assurance of quality and longevity.

5. Due to the drive to rationalise websites and also to separate presentation of data from its location, UK public sector URIs will be based around the **data.gov.uk** domain, split by sectors as sub-domains. When looking up a URI, the data.gov.uk servers either provide the response themselves, or DNS is used to redirect enquiries to the appropriate department or agency server.

6. A sector is NOT a department name. Sectors should be understandable by the public, rather than reflecting how government is currently organised. New departments taking over all or part of a sector are required to maintain the URI sets. The community of practice will provide further information about the use of sectors which are likely to be aligned to other initiatives such as Directgov.

7. Using 'education' as an example sector for a URI set promoted for re-use, gives:

   http://education.data.gov.uk

   This:
   - Shows that the set is a part of the education sector
   - Puts it in the data.gov.uk collection of UK public sector URIs promoted for re-use
   - Can be redirected using DNS to a departmental server for the content
   - Is from the data.gov.uk domain and therefore not confused with a presentation website

The path structure for URIs

General

1. The path structure of a URI may contain elements to:

Identify the set concept	**Concept** A word or string to capture the essence of the real-world 'Thing' that the set names. e.g. school • Lower case • Words separated by hyphens • Singular (e.g. 'school', not 'schools')

Identify an individual instance of a real-world 'Thing'	**Reference** A string that is used by the set publisher to identify an individual instance of concept. The reference should match the way that it is used in normal use. In some circumstances, a name may be appropriate as the Reference, e.g. 'England'. Where the name may change, or becomes overly verbose, a code may be more appropriate.

2. Examples of concept/reference pairs:
   - road/M5
   - school/123

3. The concept/reference construct may be repeated as necessary, for example:

   - road/M5/junction/24
   - school/123/class/5

4. Other components of a URI path may:

Identify the type of URI	URI type, for example one of: • id – Identifier URI • doc – Document URI, Representation URI • def – Ontology URI • set – Set URI
Identify a file format	File-extension to indicate the format of a document that will be returned. doc.{ext} e.g. doc.rdf, doc.html

The path structure for each type of URI

5. Machines should not deconstruct the path of a URI to discover other related types. However, for consistency and readability, the following can be used as a guide:

Identifier URI	Contains the string 'id' to show that it is an Identifier URI. Contains pairs of Concept/Reference
Document URI	The same as its associated Identifier URI, but replacing the 'id' string with 'doc', or removing it completely
Representation URI	The same as its associated Document URI but with a file extension
List URI	The same as a Document URI with the Reference of the final Concept/Reference pair missing. The meaning being to provide a list of URIs for that concept
Set URI	Contains the string 'scheme' and the name of the concept
Ontology URI	{domain}/def/{concept}

Looking up a URI

General

1. A URI may be used to link data without ever having to look it up. Looking up a URI is the process of resolving it (sometimes known as dereferencing). As this guidance prescribes HTTP as the protocol for UK public sector URIs, resolving a URI is via an HTTP GET.

2. The purpose of resolving a URI is to:
   - Gain greater definition of the resource that it names
   - Discover other related resources
   - Gain some basic information which a typical enquirer is likely to find useful

The client requesting the HTTP GET may be a machine or a human.

3. Referring to Figure 1, the consumer should not rely on syntactic manipulation of URIs to locate the other associated set component. The publisher should provide the ability to look-up between the components for example, given just the URI of an instance of a concept (say a particular school), to find the corresponding ontology document; thence locate information about the set which may include a catalogue of all the known concept instances. The blue arrows in the figure represent the dereferencing of a web reference taking into account any further redirection to return with a representation of the data.

4. Some of the issues to be considered when providing mechanisms to resolve a URI are given below. The community of practice will provide greater definition of the protocols/headers/status codes etc used when resolving each type of URI.

Resolving Identifier URIs

5. In a linked data architecture, it must be possible to resolve the URI for a real-world 'Thing' to a Document that contains information about that thing. In other words, it must be possible to resolve an Identifier URI to a Document URI.

6. There are two patterns for resolving Identifier URIs to provide a Document URI, as described in *Cool URIs for the Semantic Web*[2]:
    - Hash URIs
    - 303 responses

7. Either pattern meets the requirements of separating Identifier URIs from Document URIs. While 303 responses provide the most flexibility, hash URIs may prove necessary when there is limited access to the server configuration. The pattern selected by the publisher has no impact on the capability required of consumers.

8. The community of practice will provide greater definition of the protocols/headers/status codes etc used for both patterns and implementation guidance to support a decision as to which to deploy.

Resolving Document URIs

9. A Document URI will resolve to the most appropriate Representation URI to provide information in a format as requested by the client. Where more than one Representation URI is available containing the same information in different formats, this may be achieved using Content Negotiation.

10. It will often be useful to version the information that is available about a particular Thing, to indicate when the information contained in a Document is valid. It is also often helpful to journey back in time to view information that was valid in the past. To do this, there need to be 'dated' Document URIs that contain information valid on, or from, a particular date.

11. The community of practice will provide greater definition of the protocols/headers/status codes etc used when resolving Document URIs, plus advice on how to support 'dated' Document URIs, and Content Negotiation.

Coping with change and the passage of time

General

1. Once created, a URI should persist unaltered.

2. The essence of a real-world 'Thing' is unlikely to change, whereas a description of it at a point in time may change. For example, the essence of the M5 motorway remains unchanged from its original conception through to its completion. If it were to be extended, it remains the M5 motorway. Consequently, a single Identifier URI for the M5 motorway may have a number of versions of its associated Document URI that might be versioned by a status, and/or date and so on.

3. Where the essence of a real-world 'Thing' naturally passes through various stages, those stages could be designed into the structure of the URI path, thus creating a separate URI for each stage. For example:

    act/1985/67/enacted

[2] *Cool URIs for the Semantic Web*: http://www.w3.org/TR/2007/WD-cooluris-20071217/

4. The URI set publisher may provide a URI alias to the current version.

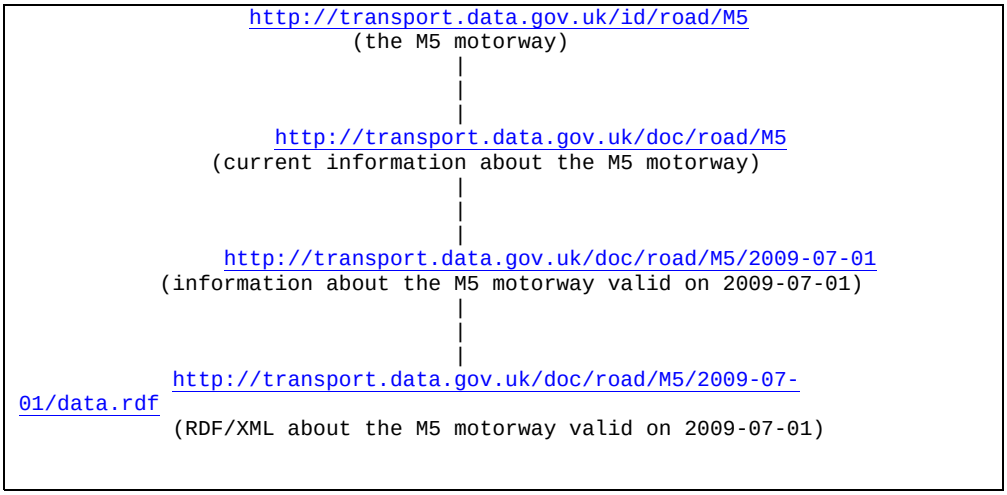
5. A URI scheme will explain how it copes with change and the passage of time and give advice to consumers about which alias URIs are appropriate to use in their statements.

6. The community of practice will provide further scenarios of how URI sets may have to cope with change and guidance as to how they can be designed to cope.

Examples of URIs within a set

URI Type	URI structure	Examples
Identifier	http://{domain}/id/{concept}/{reference} or http://{domain}/{concept}/{reference}#id	http://education.data.gov.uk/id/school/78 http://education.data.gov.uk/school/78#id http://transport.data.gov.uk/id/road/M5/junction/24
Document	http://{domain}/doc/{concept}/{reference}	http://education.data.gov.uk/doc/school/78
Representation	http://{domain}/doc/{concept}/{reference}/{doc.file-extension}	http://education.data.gov.uk/doc/school/78/doc.rdf
Definition of the scheme concept	http://{domain}/def/{concept}	http://education.data.gov.uk/def/school
List of scheme identifiers	http://{domain}/doc/{concept}	http://education.data.gov.uk/doc/school
Set	http://{domain}/set/{concept}	http://education.data.gov.uk/set/school

Example of how URIs within a set could be resolved



Machine- and human-readable formats

General

1. When a URI is resolved and a document returned a variety of formats may be used. The following is guidance as to what format to use in what circumstance.

Formats for Representation URIs

2. Documents may be available in multiple formats. Each possible representation of the Document should have a distinct representation URI. For example:

    - RDF/XML at http://transport.data.gov.uk/doc/road/M5/junction/24/doc.rdf
    - Turtle at http://transport.data.gov.uk/doc/road/M5/junction/24/doc.ttl
    - HTML at http://transport.data.gov.uk/doc/road/M5/junction/24/doc.html
    - JSON at http://transport.data.gov.uk/doc/road/M5/junction/24/doc.json

3. Having distinct representation URIs for each possible representation makes it possible to access a particular representation without changing the headers that are sent with a request, which is useful for certain clients. Each document should also include pointers to the other available formats, and lists of resources should also include lists of possible representations. For example:

    - In HTML, use the <link> element with a suitable type attribute
    - In RDF, use a dct:hasFormat property
    - In Atom, use the <atom:link> element with a suitable type attribute

4. At least one representation must be machine-readable in a way that enables the construction of an RDF graph. It is recommended that this representation is one of:

    - RDF/XML[3] (preferred, as this is supported by the largest number of tools)
    - XHTML with a GRDDL[4] transformation
    - XHTML with embedded RDFa[5]

5. It is also acceptable to provide the following formats for the construction of an RDF graph:

    - Turtle[6]
    - N3[7]

6. Other formats are also useful. In particular, it is useful to provide other machine-readable formats such as:

    - JSON
    - CSV

7. It is also useful to provide human-readable versions of the documents in HTML, if appropriate.

8. The community of practice will provide further examples to illustrate how information can be returned from looking up a Representation URI.

Governance arrangements

General

1. For these design considerations to be effective and credible, governance will need to be established to manage:

    - Maintaining this design
    - Definition of sectors
    - Allocation of lead departments/agencies for sectors
    - Engaging with stakeholders for a sector
    - Determining and prioritising URI sets within each sector
    - Avoiding naming collisions

[3] http://www.w3.org/TR/rdf-syntax-grammar/
[4] http://www.w3.org/TR/grddl/
[5] http://www.w3.org/TR/rdfa-syntax/
[6] http://www.w3.org/TeamSubmission/turtle/
[7] http://www.w3.org/DesignIssues/Notation3

- Accreditation of URI sets to be rooted at data.gov.uk
- Commitment to quality and longevity of URI sets
- Infrastructure

2. Potential publishers of URI sets may find that *they* incur cost and risk, whilst the benefits are realised elsewhere in government, or to the common good. Such obstacles should be explored, leading to proposals that make it an attractive proposition.

Appendix 1: Contributors

1. The CTO Council is grateful for participation from those representing organisations including: BBC; Ordnance Survey; Office of Public Sector Information; The London Gazette; The Stationery Office; University of Southampton; Department for Transport; Department for Communities and Local Government; Department for Children, Schools and Families; Department for Environment, Food and Rural Affairs; UK Location Programme; National Policing Improvement Agency; esd-Toolkit (Local Government); the Local e-Government Standard Body (LeGSB); Cabinet Office; Fujitsu; Hewlett Packard.

© Crown copyright 2009

The text in the document (excluding the Royal Arms and department logos) may be reproduced free of charge in any format or medium providing that it is reproduced accurately and not used in a misleading context. The material must be acknowledged as Crown Copyright and the title of the document specified.

Any enquiries relating to the copyright in this document should be addressed to The Licensing Division, HMSO, St Clements House, 2–16 Colegate, Norwich NR3 1BQ.

Fax: 01603 723000 or email: licensing@cabinet-office.x.gsi.gov.uk