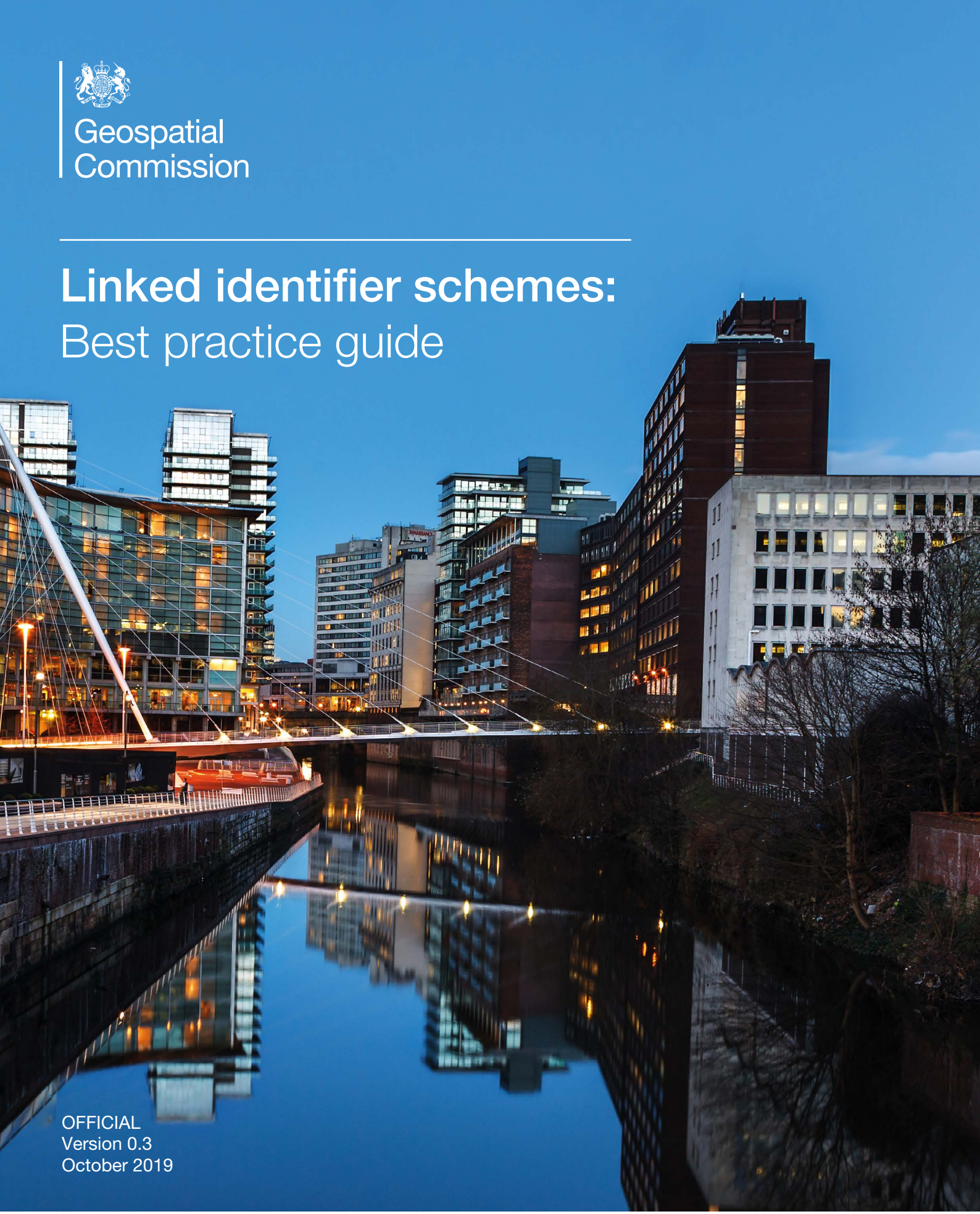
Geospatial Commission

Linked identifier schemes:
Best practice guide



OFFICIAL
Version 0.3
October 2019

The Coal Authority

HM Land Registry

Valuation Office Agency

UK Hydrographic Office

British Geological Survey
Expert | Impartial | Innovative

Ordnance Survey

Authors
This report was written by the Geo6 partner bodies, for whom the main authors are:

Organisation	Authors
British Geological Survey	Russell S. Lawley
Coal Authority	Graham Martin
HM Land Registry	Andrew Dickens
Ordnance Survey	Jack Harrison, Jonathan Simmons
UK Hydrographic Office	Nicholas Mort, Tamsin Vickery
Valuation Office Agency	Hugh Pastoll

Front Cover Image
Credit Shutterstock: SAKhan Photography

For any queries contact **geospatialcommission@cabinetoffice.gov.uk**

Contents

Introduction

In October 2018 the Geospatial Commission invested £5 million to help unlock the value of geospatial data held by its six expert Partner Bodies – the British Geological Survey, Coal Authority, HM Land Registry, Ordnance Survey, UK Hydrographic Office and the Valuation Office Agency. Our Partner Bodies, 'the Geo6', are responsible for some of the high-value geospatial datasets in the UK that underpin many of the most valuable use cases identified in the Digital Land Review.

Four projects were launched to help improve the data and unlock value and address the challenges set out in the DLR:

- making the data held by the Geo6 more easily discoverable
- simplifying the licensing landscape across the Geo6
- identifying ways of linking data from different agencies
- understanding how third-party and crowdsourced data is and could be used by the Geo6

The benefits of publishing data with persistent, unique identifiers are well understood. Within UK government, the Government Digital Service presents the challenge clearly, stating that: **"Data reusers want to be able to identify things, such as schools or companies, using identifiers that continue to mean the same thing over time. This means that data reusers can easily understand and combine data about those things from different sources."** There are also efforts under way on an international scale, with initiatives such as INSPIRE (Infrastructure for Spatial Information in the European Community) requiring unique identifiers as part of its standards.

This document builds on this work by providing practical guidance on how identifiers can be designed, created and managed to make it easy for users to understand and combine data from different sources. The primary focus of this work is the geospatial data held by the Geo6 partner bodies: the British Geological Survey, Coal Authority, HM Land Registry, Ordnance Survey, UK Hydrographic Office and Valuation Office Agency. However, the guidance is applicable to all data publishers.

When specifying a new identifier scheme, there are many design considerations. Some of these have a clear recommendation for best practice, while others present a choice between several viable options.

The recommendations in Section 1 are about designing identifier schemes that promote the efficient, reliable linking of identifiers. Making consistent design choices makes it easier for users to combine data from the Geo6 organisations.

Section 2 contains recommendations that do not directly relate to the design of identifier schemes but are complementary to linking identifiers. They should be considered alongside the recommendations in Section 1.

Terms

The following terms and definitions used in this document are provided for your reference:

- Entity – something that has separate and distinct existence and objective or conceptual reality (ISO 19119:2016). It can be a real-world phenomenon like a tree or house, or conceptual like a local authority or company.
- Identifier scheme – a logical definition of the form, update rules and metadata of identifiers that are used to identify unique representations in a dataset.
- Link – two representations that are related together by their identifiers. A link can be used to share some or all attribution between the two representations, depending on their relationship.
- Reference data – a list of data that defines permissible values in other data, such as administrative area names or classification schemes. Reference data is often defined by standards organisations, for example country codes as defined in ISO 3166-1.
- Register – a set of files containing identifiers assigned to items with descriptions of the associated items (ISO 19135-1:2015). Registers are broadly defined as lists of reference data, and are often created by a formal or official authority.
- Representations – a description or portrayal of an entity in data. For example, a row in a dataset that describes a local authority. Representations are often referred to as ‘features’ in geographic data (ISO 19101-1:2014).

Designing identifier schemes

1. Assign and use unique identifiers

Where applicable, assign unique identifiers to representations in the dataset.

Why it's important

Identifiers are labels that are assigned to representations in a dataset. Publishing your data with unique identifiers is fundamental to support the creation of links to other data. Without identifiers, it is impossible to create any kind of enduring link. Without unique identifiers, it is impossible to reliably link to specific representations in the dataset.

What it means

You should:

- assign identifiers to representations in your datasets
- ensure that identifiers are unique, at least within the scope of the dataset

2. Ensure identifier assignments are fixed

When assigning an identifier to a representation, make it immutable – unchanging over time.

Why it's important

Fixed assignments cannot be modified or reassigned once a link is made between identifier and representation.

When you ensure the identifiers assigned to representations in your dataset are fixed, any links made to those identifiers from other datasets can persist. Without this persistence, these links can quickly become invalid, or even misleading, and fall out of sync with other datasets. The result is that you lose the thread between representations common to the datasets. Links will have to be recreated every time your dataset is modified. This is especially true if identifiers are recycled over time where they are reassigned to new, different representations in your dataset.

The link between identifier and representation is only true if the representation does not change. If a change occurs, the link will need to be re-evaluated and you will need to apply best practice to maximise identifier traceability over time.

What it means

You should:

- ensure that identifiers are not modified after their initial assignment
- ensure that existing identifiers are not reassigned to new representations in your dataset

3. Maximise identifier traceability over time

Design maintenance rules so that when a change to a representation occurs, identifier assignments either persist or are traceable.

Why it's important

When a representation is changed in your dataset, often as a result of a real-world change to an entity, you will create rules that set out how it's updated or deleted. When designing these change rules, you should ensure the link between the representation and its assigned identifier persists where possible.

If the change does not fundamentally alter the representation, there is probably no need to create a new identifier for it. For example, a minor modification, such as changing the roof of an existing property, should not require a new identifier for that property. If, however, a significant change to an entity occurs, for example a house being split into two flats, it may be appropriate to retire the identifier and potentially create newly identified representations.

In situations where an identifier is retired, it is important, where possible, to allow users to trace the identifier to any other identifiers that have replaced it. In the house example above, this would mean that a link is maintained between the 'house' identifier and the two new 'flat' identifiers. Doing so makes it easier for correlations to be made between the old identifier with those that replaced it. This lets users share attribution – for example, the access road, roof type and building age.

What it means

You should:

- determine explicit rules on how to respond to changes to your dataset, such as when to retire or replace a representation
- ensure that these rules are as simple and clear as possible
- make these rules available to your users

4. Prohibit the encoding of mutable information into identifiers

Information that is liable to change, such as version numbers, should not be encoded into an identifier.

Why it's important

Sometimes it can be useful to encode information in an identifier. For example, if you included a name as part of your identifier, you could use the identifier itself to find a representation in your dataset by name.

The risk with doing this is when information included as part of an identifier changes. You will be forced to choose between either modifying the identifier or letting it persist with potentially misleading information encoded into it.

Additionally, if personal or private data is encoded into an identifier such as a property address, it may make the identifier itself personal data. This could potentially restrict the way that other datasets could link to it, by prohibiting the inclusion of the identifier in other datasets.

While not part of the identifier, abbreviations for namespaces used as affixes should also avoid encoding mutable information. This is because the preferred presentation of an identifier often includes this affix, which essentially results in the same risks as encoding that information directly into the identifier. It is sometimes necessary to include an organisation name in the namespace if the namespace needs to resolve to a web address (URL), but this mutable information should not be used in the abbreviation. The abbreviation should relate to the identifier scheme or dataset name, not the custodian of the dataset. For example, the namespace http://data.ordnancesurvey.co.uk/ontology/admingeo/ should have an abbreviation like 'admingeo' rather than 'os-admin'.

What it means

You should:

- prevent the inclusion of mutable information into identifiers
- ensure that namespaces do not encode mutable information

5. Define a single, preferred presentation of identifiers

Identifiers should have consistent presentation in published datasets. Where multiple versions of an identifier must exist, state clearly which presentation is preferred.

Why it's important

Having a consistent presentation of an identifier makes it easier to combine information from different datasets. Sometimes it is necessary to have multiple presentations of your identifier in use, for example, with or without a prefix, or as part of an HTTP URI. However, if it is unclear which is the preferred presentation, it may introduce an additional step to the process, as the identifiers must be transformed to use them together. If the process of transformation is unclear or inconsistent, this effort may be significant.

For example, a universally unique identifier (UUID) has a number of different potential encodings.

These include:

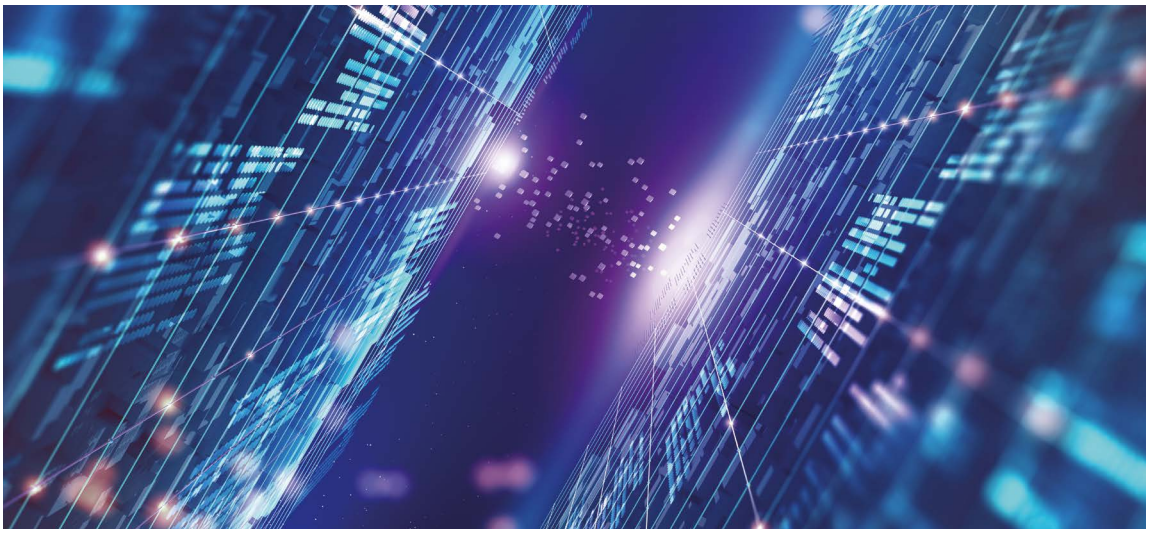
- 00112233440011 22AAAABBB-BCCCCDDDD
- 00112233-4400-1122-AAAA-BBB-BCCCCDDDD
- {00112233-4400-1122-AAAA-BBB-BCCCCDDDD }
- 00 11 22 33 44 0011 22 AA AA BB BB CC CC DD DD
- data.gov.uk/ 00112233440011 22AAAABBB-BCCCCDDDD

All of these versions (and more) could appear across different datasets as links. This would require a user to transform between the different versions before they can use the link. If, instead, a preferred presentation is specified, this effort is reduced or removed.

What it means

You should:

- where possible, publish datasets with a single presentation of an identifier
- where multiple presentations are required, state clearly which presentation is preferred for different contexts
- provide clear guidance on how to transform the identifier between different presentations, including syntax representation in a machine-readable language such as regex



Credit Shutterstock: Yurchanka Siarhei

6. Make identifiers easy to publish on the web

Ensure that identifiers only use unreserved characters that do not have a special purpose in URL syntax to make it easier to publish them on the web.

Why it's important

If you use certain characters in your identifiers, such as a forward slash (/), you cannot include the identifier in a URL without modifying it. This introduces an additional step in the publication process and creates an additional representation of your identifier to manage.

Unreserved characters are a set of 'safe' characters that do not have a special purpose in URL syntax – letters, numbers and a limited set of punctuation marks. If you ensure that your identifier only contains unreserved characters you can use it without modification, avoiding these issues.

What it means

You should:

- ensure identifiers only contain unreserved characters – the formal definition of unreserved characters that are recommended is described in the RFC 3986 Uniform Resource Identifier (URI): Generic Syntax document

7. Provide metadata on the life cycle of representations

Support the understanding and management of changes that occur within a dataset by providing metadata about the life cycle of each representation.

Why it's important

It's easier to understand and manage datasets when you know what changes to representations have occurred and why.

Including a life cycle status value (e.g. 'active', 'inactive') allows you to publish a complete set of representations in a dataset. This is because it provides a way to differentiate between the 'live' view of the dataset and the full set that includes identifiers that are no longer active. If the complete set of representations is available, a link made to one of your identifiers will persist indefinitely. Without a life cycle status, you would typically only include active identifiers in your dataset release, which would make links to inactive identifiers from other datasets unresolvable.

You should use life cycle metadata about updates, such as version numbers, to evaluate the validity of links made to your dataset. A link between two representations in different datasets is only guaranteed to be valid if neither representation changes. Where a change occurs, the link needs to be re-evaluated. Life cycle metadata alerts users to the need for an evaluation, ensuring they do not unknowingly use invalidated links. It is also useful to provide a codified, well-described reason for the change so that a user can determine whether the change invalidates any links they have made. This would save the effort of recalculating them in some cases.

Users can also use life cycle metadata to assess data quality when creating links; for example, to check that the currency of the data is sufficient for their needs.

What it means

You should include the following information for each representation in your dataset:

- the date and time that the representation was added to the dataset
- the version of the representation
- the date and time that the representation was last updated
- a codified, well-described reason for the update
- the life cycle status of the representation

8. Make published identifiers free of licensing restrictions

Your identifiers should be able to be published freely, without requiring a licence for the dataset.

Why it's important

If the dataset will not be released under an open licence, it is beneficial to make the identifier itself free of licensing restrictions. Doing so permits the reuse of the identifier by other data publishers, encouraging them to create links to your identifier or even adopt your identifier scheme for their own data. This in turn enriches your dataset, allowing users to easily combine it with other, related datasets with very little effort or difficulty.

To realise these benefits, it is also useful to make the life cycle metadata free of licensing restrictions. This further encourages adoption by other data publishers, as linked identifiers can be validated freely using this metadata. Without it, it's impossible to know whether the linked identifier was still active, or whether it had been updated since the link was created – reducing its utility.

What it means

You should:

- release a policy statement setting out the permitted use of the identifier scheme and associated life cycle metadata. Where possible, this permitted use should be free of any licensing restrictions. Where this approach is not possible, the statement should clearly explain the reasons for any restrictions.

9. Comprehensively document the identifier scheme

Make the design decisions of your identifier scheme explicit by clearly documenting them.

Why it's important

Many of the decisions made when designing your identifier scheme will not be apparent to a user simply by looking at the dataset. For example, you cannot tell by looking at a single release of your dataset whether or not you have prohibited the reassignment of identifiers. This means that most of the benefits of the other points in this guide are reliant on clear, comprehensive documentation.

What it means

You should:

- provide documentation of your identifier scheme, using the standard template provided alongside this document

Credit Pexels: Pierre Blache

Other considerations

10. Use common reference data

Assign your reference data from common, authoritative sources such as GOV.UK Registers.

Why it's important

Reference data, such as administrative area or classification codes, is often assigned to dataset representations to describe them.

There is often a great deal of variation in the presentation of reference data, including difference in spelling, abbreviation and language. This can make it difficult to create links between two datasets that use different reference data.

However, if two datasets have assigned reference data from the same register, it is much easier for a link to be made between them. If a user is preparing aggregate statistics, they may be able to link the datasets using the reference data alone. For example, it may be appropriate to report statistics at local authority level, and if both datasets use the same local authority names, it is a simple process to link them. If the user needs to link the data at a more granular level, common reference data can make that process more efficient by first making a link using reference data and then using other attribution for a more granular link.

To best realise the benefit of this recommendation, it is important to add the identifier (or 'key') from the register, not just the reference data itself (e.g. 'E10000001', as well as 'Bedfordshire'), to facilitate efficient linking through keys.

What it means

You should:

- assign reference data from common, authoritative sources where possible
- where possible, include the identifier for the assigned reference data in addition to the data value itself

The principle source of common reference data for UK public sector datasets is the GOV.UK Registers service. It provides structured datasets of government information, each maintained by a subject matter expert from the most relevant government organisation.

11. Provide supporting services for identifiers

Allow users of your identifiers to access metadata through supporting services.

Why it's important

If you make the metadata of your identifiers accessible freely through a service, it will be easier and more appealing for other publishers to include your identifiers in their datasets.

Whether through creating and maintaining links to your identifier or even adopting your identifier scheme for their own dataset, this reduces the amount of matching and correlation that needs to be performed by end users.

These services are also beneficial to users to validate and check the current status of your identifiers – whether the identifier appears in your dataset or as a link from another dataset.

What it means

You should implement a metadata dereferencing service that provides, for a given identifier in your dataset:

- the custodian and dataset name associated with the identifier
- the version of the representation associated with the identifier
- the date and time that the associated representation was last updated
- a codified reason for the update
- the life cycle status of the associated representation

12. Comprehensively document links to other datasets

Where you publish links from identifiers in your dataset to those in other datasets, support the understanding and evaluation of those links by clearly documenting them.

Why it's important

To support your users, you may decide to calculate and publish a link between an identifier in your dataset and an identifier in another dataset. However, links made between different datasets are usually incomplete and imperfect. The process typically involves a number of design decisions such as the method you use to create the links and a confidence measure. For example, if your dataset contains property information you may wish to append some sort of address identifier to the representations to make it easier for your users to work with. Making this link would typically require you to match the text of the addresses, and you'd need to decide how strict that matching process is. For example, do you require the addresses to be identical, or are small changes in capitalisation or abbreviations acceptable?

If you only publish the linked identifiers, it can be hard for users to understand if the way the link was made is suitable for their use case. Often this means that users assume that the link is authoritative and complete, when in most cases it is inferred and partial. By providing sufficient documentation on the link, they can evaluate it for themselves.

What it means

You should:

- provide documentation for each link made to other datasets, using the standard template provided alongside this document