Guidelines for BS 7666:2006

Version 1

January 2007

Section 1. Introduction to BS 7666

Section 2. How to create a gazetteer of a new type of geographic object

Section 3. Quality assessment and reporting

Section 4. How to create a ‘national’ gazetteer

Preface

These Guidelines are intended for use with BS 7666: 2006 *Spatial datasets for geographical referencing*. They complement the Standard with more detailed explanation of the content and a general approach to creation of gazetteers. They are not specific to any particular implementation, for which it is expected that specific data specifications and capture and maintenance rules will be produced.

The Guidelines are aimed at:

- gazetteer owners – those with ultimate responsibility for the gazetteer;
- gazetteer custodians – those responsible for the creation, maintenance and quality of gazetteers;
- suppliers of gazetteer software;
- those developing and implementing gazetteer systems
- providers of gazetteer data;
- others who are responsible for aspects of the quality management of gazetteers.

The Guidelines are currently in four Sections:

1. Introduction to BS 7666;
2. How to create a gazetteer of a new type of geographic object;
3. Quality assessment and reporting;
4. How to create a national gazetteer.

Further Sections will cover specific implementation issues:

- How to create a street gazetteer;
- How to create a land and property gazetteer;
- How to create a delivery point gazetteer.

No guidelines for public rights of way which form an informative annex to Part 1 of the Standard are planned at present.

These Guidelines have been produced under the auspices of BSI IST/36 geographic information who are responsible for BS 7666. They were written by Rob Walker and Les Rackham working under the guidance of a Steering Group comprising representatives of major stakeholders in the Standard. The work is sponsored by the Department for Communities and Local Government (DCLG), Ordnance Survey, Office for National Statistics and Mayrise Ltd.

Contents

Section 1. Introduction to BS 7666

This Section provides a general introduction to all Sections of the Guidelines. It gives a general introduction to the subject of gazetteers and addressing. It also provides an introduction to standards, describes what BS 7666 is, identifies the main changes made in the 2006 edition and provides an overview of the other Sections of the Guidelines. A Glossary of Terms, a list of abbreviations used in the Guidelines and an explanation of the UML data modelling convention used in the Standard are included in Annexes.

1. General introduction to all parts of the Guidelines

This document is an introduction to a set of guidelines to accompany the British Standard BS 7666: 2006 *Spatial datasets for geographical referencing*. The 2006 edition is in four parts:

- Part 0: *General model for gazetteers and spatial referencing*;
- Part 1: *Specification for a street gazetteer*;
- Part 2: *Specification for a land and property gazetteer*;
- Part 5: *Specification for a delivery point gazetteer*.

Part 3 *Specification for addresses* and Part 4 *Specification for recording public rights of way* of the 2002 edition have been withdrawn, their content having been subsumed in the other parts.

BS 7666, particularly Parts 1 and 2 has been very widely used by the local authority community who are creating local street gazetteers and local land and property gazetteers. These in turn are being merged to form the National Street Gazetteer (NSG) and the National Land and Property Gazetteer (NLPG). Feedback from these activities has influenced the revision of the Standard and identified the need for further explanation and guidance.

Standards are not easy things to read or understand. Ideally, they express unambiguous requirements and do not include lengthy explanations. The emphasis is on precision in the use of language such as to minimise ambiguity. The intention is that standards should be used by those with knowledge of the subject area to build implementations that suit their specific applications.

This approach can work well where the objects in scope of the standardisation are susceptible to clear and close definition, or where there is a large and agreed body of theory and practice. Unfortunately, this is not the case with geographic information where the objects (e.g. buildings, paths, railways, rivers, woodlands) may be tangible but are not that well-defined or identified, and the theory and practice is still evolving.

Some form of additional guidance, which complements BS 7666, and which is written in a more accessible form is needed. It can be assumed that the audience will be familiar with the objects in question (e.g. streets or land and property) but not necessarily with the requirements set by the Standard or the economy of language.

To make the Standard more accessible and usable, the Guidelines aim to:

- provide an introduction to all parts of the Standard;
- interpret the requirement clauses in the Standard;
- provide general guidance in the implementation of the Standard;
- provide illustrations and examples.

In terms of scope, the Guidelines cover:

- the underlying concepts, the purpose of the Standard and the applicability;
- the approach to creating a gazetteer of any type of geographic object in conformance to the general model in Part 0 of the Standard;
- the general approach to the creation of gazetteers of streets, land and property and delivery points;
- quality assessment and reporting;
- the creation of national gazetteers.

It is intended that the Guidelines should be for general use and independent of any particular application. Specifications for these applications will need to be developed by the bodies responsible for their implementation but within, it is hoped, the framework provided by these Guidelines. For example documentation is being developed to support the National Street Gazetteer (NSG), National Land and Property Gazetteer (NLPG) and DNA-Scotland (Definitive National Addressing – Scotland) implementations of the 2006 edition of the Standard. It is expected that such implementations will have their own data specifications and capture and maintenance rules based upon user requirements. These concepts are illustrated in Figure 1.

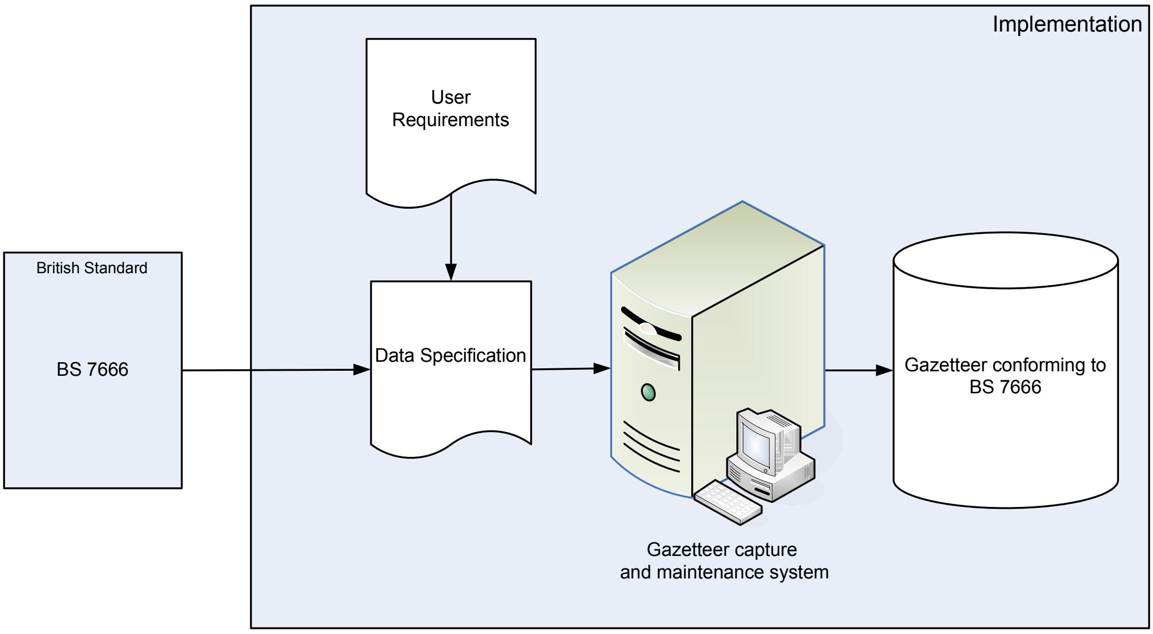

Figure 1: The relationship between the Standard and a gazetteer implementation

These Guidelines should be read in conjunction with the Standard. The Standard provides the authoritative statement of what is required, and the Guidelines are recommendations on how to fulfil those requirements. Throughout this document, the term ‘the Standard’ refers to all parts of BS 7666.

Examples or explanations of certain terms and more detailed concepts are enclosed in boxes. This is to aid those readers who are less familiar with the Standard and the concepts while enabling others to pass over them. There is also a glossary of terms and abbreviations in Annex A.

2. Standards

2.1 What is a standard?

The term ‘standard’ is often associated with a perceived quality level e.g. some product or service is said to reach “a high standard”.

In the context of information and communications technologies (ICT) and specifically BS 7666, ‘standard’ has a rather different meaning. The ‘standard’ provides a norm, model or specification which should be conformed to when implementing systems or structuring information. By so doing, interoperability and information exchange are enabled. Actual quality levels are not specified in the standard, these are left to the implementation so that levels can be established taking into account user requirements.

Characteristic of these sorts of ICT standards, including BS 7666, is a specification that:

- provides rules, guidelines or characteristics for activities or their results – i.e. it expresses a requirement in a formal manner;
- is aimed at the achievement of the optimum degree of order in a given context - i.e. it must have some practical application;
- is established by consensus – it is not ordained from on high;
- is approved by a recognised body, such as BSI, the International Standards Organisation (ISO), or the Open Geospatial Consortium (OGC);
- is for common and repeated use – standards have no value unless widely applied.

2.2 How requirements are expressed in standards

When requirements are expressed in a standard, particular language is used to indicate whether this is mandatory or optional. In the case of British Standards, if the requirement is mandatory then the verb ‘shall’ is used but if it is optional then ‘may’ is used. For example, “the gazetteer *shall* have the following metadata elements” means that these must be present to conform to the standard but “the location instance *may* be related to other location instances” means that the location instance can be related to another instance but it does not have to be to conform. The term ‘should’ is used where making a recommendation.

2.3 Conformance to Standards

Standards by themselves only become enforceable when mandated, for example by government regulation. Adoption is usually voluntary and brought about because it is commercially beneficial or in the public good. Organisations often claim conformance or compliance as an indication of following best practice. However, this is often a statement of intent rather than a verifiable fact. Many standards, including BS 7666, now contain a "conformance clause" or "statement of conformity", which states what the implementer must do to demonstrate conformance with the standard. This will often take the form of a set of tests that should be passed or can be verified to have been passed.

2.4 The value of standards for geographic data

The benefits of standards vary with the area of application, but generally include the following:

- increased user confidence that their needs are being met;
- compatibility between similar implementations, enabling data sharing between users, interoperability between applications and creation of national datasets;
- increased data integrity;
- greater understanding of products;
- reduced production costs and reduction of data duplication.

3. Gazetteers and addressing

3.1 Geographic objects

BS 7666 is concerned with gazetteers, lists of geographic objects of a given type with their locations which provide information identifying and describing where they are in the real world. Any type of object that is fixed in position and can be consistently identified, recognised and described as occupying a specific place in the real world can be regarded as a **geographic object**. It can range in size from a lamp post to an administrative area or even a complete country depending on the application.

Examples of geographic objects
Commonly used examples of geographic objects are streets and occupied properties. Other examples are countries, towns, localities and postcode areas.

These geographic objects are used to define a location which is turn is used to reference other data. The purpose of gazetteers is to allow the consistent linking of information to locations and by so doing enable the cross-referencing of all information relating to those locations.

Example of geographic objects used as locations to reference other data
Residential properties, referenced by their address, are used to reference demographic data.

3.2 Spatial references

A location is given a **spatial reference**. In the context of BS 7666, this spatial reference takes the form of a description of a real-world place which can then be used to reference other information. Each location needs to be uniquely identified, by a name (e.g. for a town) or a code (e.g. a postcode). The name may not in itself be unique (e.g. there are several occurrences of 'Newport' around the world), and where this occurs, further information is need to distinguish between them (e.g. 'Newport, Isle of Wight').

In building up a unique spatial reference, other locations such as locality, town, county have to be used which are themselves geographic objects. These other locations used in referencing the location in the gazetteer are termed **spatial units**. Spatial units can reference other spatial units (e.g. 'Ashford, Kent') and typically form part of a hierarchy of spatial units. The way that spatial references are applied to locations is specified in the **spatial referencing system** which details the types of spatial units and their relationships.

The most common example of a spatial reference is an **address.** This locates and identifies the object using a standard set of spatial units. Instances of these units are identifiable in the real world by their name. A geographic object that is capable of referencing in this way is termed an '**addressable object**'. A common form of an address is the postal address. This is defined by Royal Mail for the purpose of delivery of mail. It applies only to buildings that receive mail deliveries, and uses a set of spatial units, such as post town, that represent the organisational structure of the delivery organisation. However, a geographic address can be produced for other types of addressable object that do not have a postal address. This latter approach is used in BS 7666.

Geographic addresses

Defined in BS 7666, these are designed to be applicable to a range of geographic objects such as land and property, not just properties that receive mail. They are based upon the following:

- object name – the number, name, description or occupant of the property;
- street reference – the name or number of the street;
- locality name – an established name for the local area;
- town name – the nearest town;
- administrative area name – this is the highest level local authority, usually a county or unitary authority.

An example of a geographic address is 'The Cenotaph, Whitehall, London'.

In addition to the descriptive spatial references described above, a geographic object also has a **position** defined by coordinates such as those derived from the National Grid of Great Britain. Recording the position of a geographic object in a gazetteer is mandatory in the Standard but is never the sole means of locating it.

The concepts of spatial referencing and gazetteers are illustrated in Figure 2.

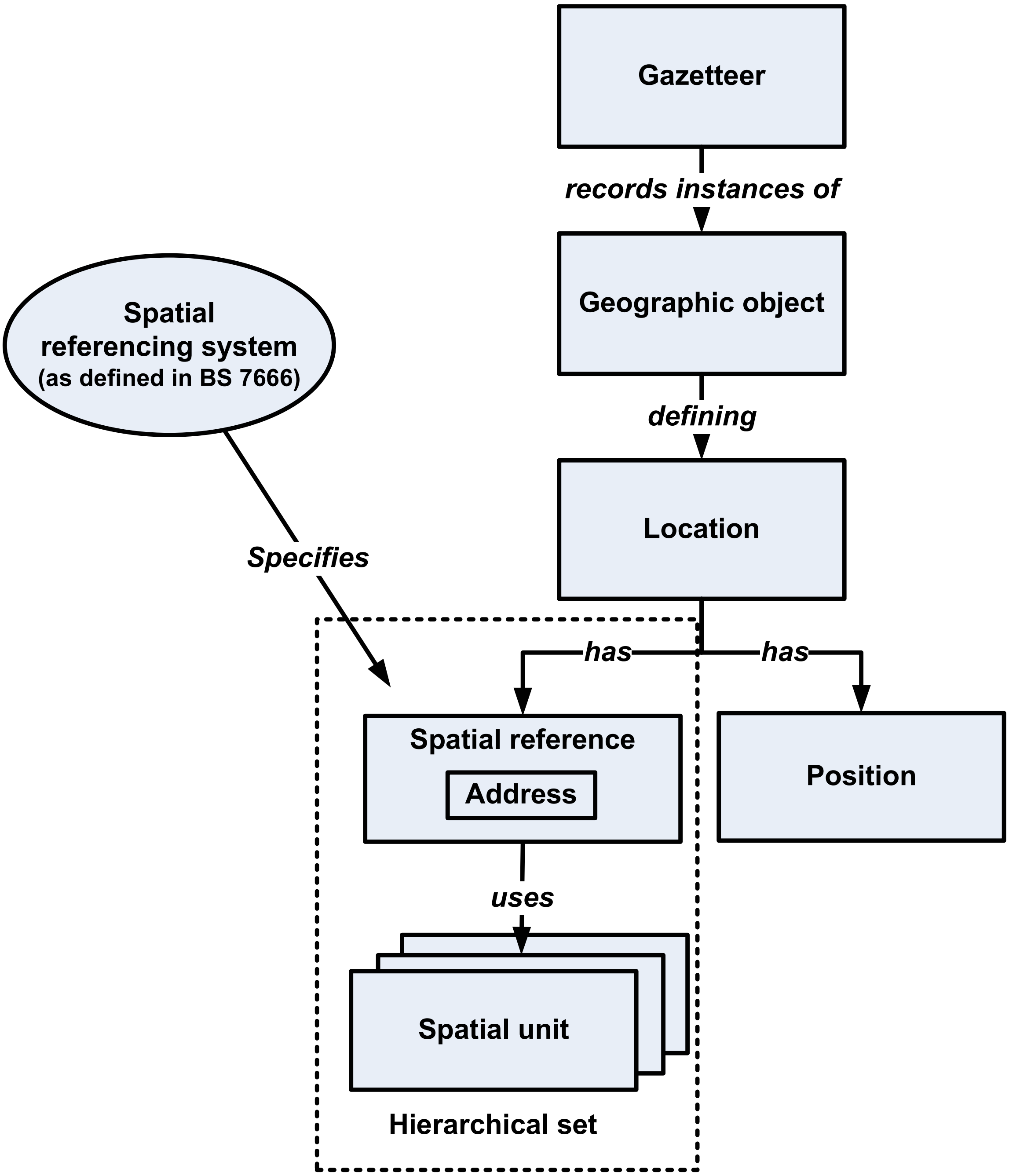

Figure 2: Concepts of spatial referencing

To make each geographic object uniquely identifiable within a gazetteer, a **unique identifier** is needed for each geographic object. This is usually in the form of a unique number which is retained by the object throughout its life-cycle from creation in the gazetteer to its demise.

4. What is BS 7666?

4.1 First edition

BS 7666 was first published over the period 1994 to 1996 in four parts. Part 1 was originally produced to support street works activities. Its main application was the production of street gazetteers in local authorities, which were then combined to form the National Street Gazetteer.

Part 2 followed to support land and property gazetteer production in local authorities. This required a mechanism for creating addresses for a range of objects, not only those that have postal addresses (i.e. receive mail deliveries), including unoccupied land, industrial premises and some public buildings, and Part 3 was produced to provide a general address structure for any such geographical objects. Part 2 was used extensively in local authorities in the creation of local land and property gazetteers, and underpins the National Land and Property Gazetteer (NLPG).

Part 4, first produced in 1996, was somewhat different from the other Parts, being not a gazetteer specification, but a specification of additional information to be recorded about Public Rights of Way (PROW), which were themselves linked to streets.

4.2 2000-2002 edition

All four parts were revised separately over the period 2000-2002. During 2003, a strategic review of the standard was carried out. This concluded that:

- there was a strong and continuing business case for the Standard;
- it had been successful in the context of the National Street Gazetteer (NSG) and the National Land and Property Gazetteer (NLPG);
- the context within which the Standard operated was changing, other related national datasets were being developed including OS MasterMap® (with its address and transport layers), new technology was allowing greater interoperability between systems and datasets, new International Standards were emerging;
- a general model of spatial referencing and gazetteers was required, for use in particular by those outside the local authority community.

It resulted in the current revision of all parts of the Standard together, and the consideration of a wider scope for the Standard.

4.3 Purpose of the Standard

The general purpose of BS 7666 is to define standard referencing systems for a range of types of geographic objects. These include not only streets and land and property, but also other classes of reference objects such as localities, towns and countries. It provides a standard way for identifying and defining these geographic objects, and of sharing and accessing information about them. Specifically, it assists the creation of local datasets or gazetteers which will in turn enable the creation of national datasets or gazetteers.

4.4 Benefits of adoption

The general benefits of adopting BS7666 are that data created by different users is produced to a common specification, so that data can be readily interchanged, and amalgamated to create ‘national’ gazetteers. This has been the basis of implementations such as the National Street Gazetteer (NSG), National Land and Property Gazetteer (NLPG) and Definitive National Addressing – Scotland. Adoption of the standard also means that common processes can be applied, making implementation more straightforward.

4.5 Perspectives on the Standard

There are different perspectives that need to be considered in any implementation of the standard:

- **Gazetteer users**: who need to know what they will find in the gazetteer, and get some idea of its quality, as well as being able to access the data;
- **Gazetteer custodians**: those responsible for creating and maintaining the gazetteer;
- **Gazetteer owners**: those who retain the intellectual property rights (IPR) to the gazetteer (but not necessarily to the IPR of every item of data in it);
- **Data suppliers**: those who supply the data, in whole or in part;
- **System suppliers**: who provide software for gazetteers and related applications.

4.6 Object names

BS 7666 standardises only the structure and form of the gazetteers. It does not standardise the content itself. Gazetteers are essentially records of geographic place names. These place names are created and controlled elsewhere, sometimes by a statutory body. Of particular importance to the Standard is street naming and numbering. This is a statutory function carried out in local authorities. This process is outside the remit of BS 7666. Many of the problems encountered in addressing relate to street naming and (property) numbering, and need to be resolved in that context. Complex gazetteer structures and records are not a sensible way of solving real-world addressing problems. There is no substitute for rational street naming and property numbering, and data standards cannot solve problems resulting from institutional failure.

4.7 International Standards

BS 7666: 2006 is based upon an International Standard ISO 19112 *Spatial referencing by geographic identifiers* that deals with gazetteers. The principles and concepts employed are the same but there are some differences in terminology. The International Standard is described in Annex C, which also explains the differences between it and BS 7666.

5. Changes to the Standard

5.1 What are the changes introduced in the 2006 revision?

The main changes are as follows.

- Introduction of a new Part 0 to provide a common structure for gazetteers of any class of geographic object.
- Harmonisation of structure, content and terminology across all parts to provide consistency with Part 0.
- Use of UML (Unified Modelling Language) for presenting data models (see Annex D);
- Addition of a requirement to provide metadata for all gazetteers.
- Addition of a facility for recording descriptive identifiers in multiple languages.
- Addition of a facility for classification of objects recorded in the gazetteer.
- Extension of facilities for cross-referencing to other datasets in all Parts of the Standard.
- Addition of a requirement for a data quality report for all gazetteers.
- BS 7666 Part 3, *Specification for addresses* is withdrawn, and its contents incorporated in Part 0 as an informative annex.
- BS 7666 Part 4, Specification for recording Public Rights of Way is withdrawn, and its contents incorporated in Part 1 as an informative annex.
- Introduction of a new Part 5 *Specification for a delivery point gazetteer*. This is applicable to a range of delivery services.
- Field lengths are no longer prescribed.
- Minor changes in the light of experience.

The main changes to Part 1 are:

- Removal of the tolerance attribute and the replacement of the spatial locator by a pair of extremity points;
- Removal of the requirement for the identifier of an elementary street unit to be the coordinates of a reference point, and replacement with a general identifier;
- Addition of an informative annex on Public Rights of Way.

The main changes to Part 2 are:

- The structure and content are made consistent with the revised Part 1;
- Addition of an identifier for a Land and Property Identifier (LPI);
- Changes to the role of provenance.

Further details of these changes can be found in the relevant sections of these Guidelines dealing with Parts 1, 2 and 5 of the Standard.

5.2 What are the general implications of the changes?

Most of the changes concern presentation and clarification. However, there are some additional requirements introduced (for example the requirement to produce metadata and quality reports[1]), and some additional enabling facilities (for example classifications and cross-referencing). Changes to the detail are small, are based upon current best practice, and are highlighted in the relevant sections of these guidelines.

6. Overview of other sections of the Guidelines

The Guidelines are currently in four Sections.

Section 2 of these Guidelines covers how to create a gazetteer of a new class of geographic object. Part 0 of BS 7666 defines a generic gazetteer model that can be used for any class of geographic object, and this section describes how to implement this. It describes the planning and initiation of such an implementation and general principles of gazetteer construction. Finally the process of creating a new part of BS 7666 is outlined.

Section 3 covers quality assessment and reporting. It introduces the principles and concepts of data quality in the context of gazetteer creation, maintenance and utilisation and provides general guidance on how to test and report on gazetteer quality in conformance with BS 7666: 2006. This includes a description of some basic data quality measures and quality evaluation procedures, some quality test methods, the outline of a data quality report and an overall process for controlling quality of a gazetteer. It is at a generic level. It will be expanded in later sections in the context of particular types of gazetteer.

Section 4 deals with how to create a ‘national’ gazetteer. It describes how to amalgamate ‘local’ gazetteers into a ‘national’ one. It identifies quality issues between the two types of gazetteer and in particular issues relating to how a national gazetteer should be maintained.

Further Sections will cover specific implementation issues relating to the specific parts of BS 7666:

- How to create a street gazetteer;
- How to create a land and property gazetteer;
- How to create a delivery point gazetteer.

[1] Quality reports were previously a requirement for Part 1 only

Annexes

Annex A. Glossary of terms

The Standard contains many technical terms, and uses some in a very specific way. This annex provides an explanation of the terms used, rather than formal definitions, which can be found in the Standard. Other terms used are also defined in the Standard.

Acceptable quality levels (AQLs)
threshold values applied to the results of testing data quality to determine whether the data meets criteria determined from a data specification or user requirements

Address
type of spatial reference in the form of the names or numbers of a sequence of spatial units used to identify and locate a geographic object, applicable to a wide range of geographic objects and not just properties that receive mail

Addressable object
geographic object that is capable of being referenced by an address

Conditional (when applied to an attribute)
value is mandatory where a stated condition is satisfied

Currency
indication of how up-to-date a gazetteer is

Current date
date at which the information in a gazetteer is considered to be current

Current state date
actual date when a geographic object came into its current state in the real world rather then when it was recorded in the gazetteer

Data specification
complete description of the data required to fully implement a system, including details of the items to be included, data structure, types, formats and definitions, classification systems, and quality levels to be achieved

End date
actual date when a geographic object ceased to exist in the real world rather than when this was recorded in the gazetteer

Entry date
date when a record was entered into the gazetteer

Gazetteer
lists of geographic objects with information identifying and describing where they are in the real world

Gazetteer custodian
person responsible for the creation, maintenance and quality of a gazetteer

Gazetteer owner
person, persons or organisation with ultimate responsibility for the gazetteer

Geographic extent
detailed description of the "footprint" of the geographic object, recorded either as a collection of one or more geographic objects or as one or more boundary polygons

Geographic object
any type of object that is fixed in position and can be consistently identified, recognised and described as occupying a specific place in the real world. This can range in size from a lamp post to an administrative area or even a complete country depending on the implementation

Lineage
history of the dataset, the sources used, the maintenance applied and the methods used in the derivation of the data and changes made since its inception

Location
identifiable geographic place

Logical consistency
degree of consistency to rules for the recording and encoding of data items within the gazetteer

Mandatory (when applied to an attribute)
value must always be provided

Metadata
data describing other data. In the case of a gazetteer, this is data about the name, scope, territory of use, owner, currency and other essential information needed to be able to use the gazetteer

Metadata date
date when the metadata was last updated

Optional (when applied to an attribute)
value may not be provided where it is not applicable

Parent-child
relationship between a geographic object (the child) which forms part of, or is dependent in some way on, another geographic object (the parent) for example a flat within a larger building, a shop unit within a shopping mall

Primary addressable object
geographic object that can be addressed without reference to any other addressable object included in the gazetteer

Primary classification
classification of geographic objects for high-level, external purposes to provide a coarse division of the objects

Quality
totality of characteristics of a product that bear on its ability to satisfy stated and implied needs. Can be more simply expressed as fitness for purpose

Quality assurance
overall planning of production processes to ensure that the product meets the required quality levels

Quality control
the way in which quality checks are carried out during the production process and those items failing quality checks are managed

Secondary addressable object
geographic object that can only be addressed by reference to a primary addressable object included in the gazetteer i.e. it is the child of the parent addressable object

Secondary classification
more detailed classification of geographic objects than the primary classification, used mainly for internal or application-specific purposes. It is not necessarily a refinement of the primary classification scheme, and may be completely independent of it

Spatial reference
description of a real-world place which can then be used to reference other information, for example an address

Spatial referencing system
specification of the way that spatial references are applied to locations and details the types of spatial units and their relationships

Spatial unit
other locations used in referencing the geographic object in a gazetteer e.g. locality, town, county, country

Start date
actual date when a location came into existence in the real world, as defined in the implementation

Unique identifier
descriptor or number applied to one and only one instance of a geographic object which is retained by the object throughout its life from creation in the gazetteer to its deletion from the gazetteer

Update date
date when a location record was last updated

Annex B. Abbreviations

AGI – Association for Geographic Information

AQL – Acceptable Quality Level

BLPU – Basic Land and Property Unit

BS – British Standard

BSI – British Standards Institution

DfT – Department for Transport

DNA-S – Definitive National Addressing - Scotland

DPC – Draft for Public Comment

ESU – elementary street unit

ETRF 89 – European Terrestrial Reference Frame 1989

IST/36 – British Standards committee for geographic information

LPI – Land and Property Identifier

NLPG – National Land and Property Gazetteer

NSG – National Street Gazetteer

OS – Ordnance Survey

ONS – Office for National Statistics

PAF – Postcode Address File

PROW – Public Right(s) of Way

RMSE – Root Mean Square Error

SE – Standard Error (of the Mean)

TOID – Topographic Identifier

UML – Unified Modelling Language

UPRN – Unique Property Reference Number

USRN – Unique Street Reference Number

Annex C. ISO 19112 Spatial referencing by geographic identifiers

ISO 19112 *Geographic Information – Spatial referencing by geographic identifiers* was published in 2003 and follows the approach of BS 7666. ISO 19112 uses the term 'geographic identifier' for a spatial reference in the form of a label or code that identifies a location. It is more generic than BS 7666. It defines a conceptual schema for spatial references based on geographic identifiers. It establishes a general model for spatial referencing using geographic identifiers, defines the components of a spatial reference system and defines the essential components of a gazetteer.

ISO 19112 has been adopted as a European Standard (EN ISO 19112) and as a British Standard (BS EN ISO 19112).

ISO 19112 provides a specification for a spatial referencing system using geographic identifiers. BS 7666-0 does not attempt to record the details of a spatial referencing system per se, only defining the individual spatial units, and the spatial referencing system used in the gazetteer is identified by name.

ISO 19112 specifies the following properties for a gazetteer, equivalent to the gazetteer metadata specified in BS 7666-0:

- Identifier – equivalent to 'name' in BS 7666-0;
- Scope – as in BS 7666-0;
- Territory of use – as in BS 7666-0;
- Custodian – as in BS 7666-0;
- Coordinate reference system – termed 'coordinate system' in BS 7666-0;
- Location type – not explicitly held in metadata in BS 7666-0, but implicit on the model through the relationship with **location**.

ISO 19112 distinguishes between the location type (the class of object) and the location instance (the individual occurrence). It specifies the following attributes for a location instance:

- Geographic identifier – termed 'identifier' in BS 7666-0;
- Temporal extent – equivalent to 'start date' in BS 7666-0, it is used in ISO 19112 to identify the version of the location;
- Alternative geographic identifier – termed 'alternative identifier' in BS 7666-0;
- Geographic extent – termed 'extent' in BS 7666-0;
- Position – as in BS 7666-0;
- Administrator as in BS 7666-0;
- Parent location instance – termed 'parent' in BS 7666-0;
- Child location instance – termed 'child' in BS 7666-0.

ISO 19112 additionally specifies a set of attributes of a location type (equivalent to metadata elements for a location in BS7666-0) as follows:

- Name – as in BS 7666-0;
- Theme – the property used as the defining characteristic of the location type, which would be included in the scope in BS 7666-0;
- Identification – the method of uniquely identifying location instance, which is not included explicitly in BS 7666-0, but is implicit from the form of the identifiers;
- Definition – in BS 7666-0, this is held as an attribute of ‘spatial unit’, as it is only required when the location is used as a spatial unit;
- Territory of use – this is held in the gazetteer metadata in BS 7666-0;
- Owner – equivalent to ‘administrator’ in BS 7666-0;
- Parent location type - called ‘ parent’ in BS 7666-0;
- Child location type – called ‘child’ in BS 7666-0.

ISO 19112 cites as an example of a spatial referencing system, a geographic address as defined in BS 7666.

Annex D. Explanation of Unified Modelling Language (UML) notation

D.1 UML diagrams

The model diagrams in the Standard use UML (Unified Modelling Language). These diagrams show the object classes, their attributes and the associations between the classes.

D.2 Object Classes

Classes are shown as boxes. These boxes are often in two parts, with the name of the class shown in the upper part, and the attribute in the lower part. Where no attributes are given, only the upper part is shown.

D.3 Attributes

Attributes are listed with their name, the minimum and maximum number of occurrences, and their data type. The name of the attribute is unique for the object class. The minimum and maximum number of attributes are given as a range in brackets. [0..] indicates that the attribute is optional. [..n] indicates that multiple values are allowed. Where no number range is given, a single attribute value is mandatory.

Attribute multiplicity and conditionality

0..1 means that a single (optional) value may be given

0..n means that multiple (optional) values may be given

1..n means that multiple values may be given, but one value must be given (mandatory)

No numbers means that a single value must be given

The data type of the attribute is shown. This is the form of the value of the attribute. In some cases, the attribute takes the form of another object class defined in BS 7666 (i.e. has a specific defined structure). These other classes may take a value from a code list or be a set of attributes. Examples of this are GeogExtent and Point.

Standard data types used in BS 7666

CharacterString: a sequence of alphanumeric characters

Integer: a whole number

Date: a date according to BS ISO 8601, in the form YYYYMMDD or YYYY-MM-DD

D.4 Associations

Associations are relationships between classes. They are indicated by lines (links) between the boxes. They may be identified by name (e.g. "aggregation") or by a role (e.g. "has"). The role is that played by the class (the source) in the association from the perspective of the other class (the target). Thus an LPI "identifies" a BLPU, whilst a BLPU is "identBy" an LPI. Also shown is the multiplicity of the association from the perspective of the other class (the target) to that class (the source).

Multiplicity and optionality of associations

0.. 1 means that the association is optional

1 means that there must be one occurrence of the association only

1..n means that there must be one occurrence of the association, but may be more

0.. n means that there may be zero, one or more occurrences of the association

Associations are usually implemented as an attribute (the role) with the target class as the data type. There are some standard types of associations that are used.

Standard types of association

◊ represents an aggregation, i.e. one class is made up of instances of another, e.g. a gazetteer is an aggregation of locations.

Δ represents a generalisation, i.e. one class is a sub-type of another, e.g. Bounding Polygon and Geographic Object are types of Geographic Extent.

These concepts are illustrated in Figure 3.

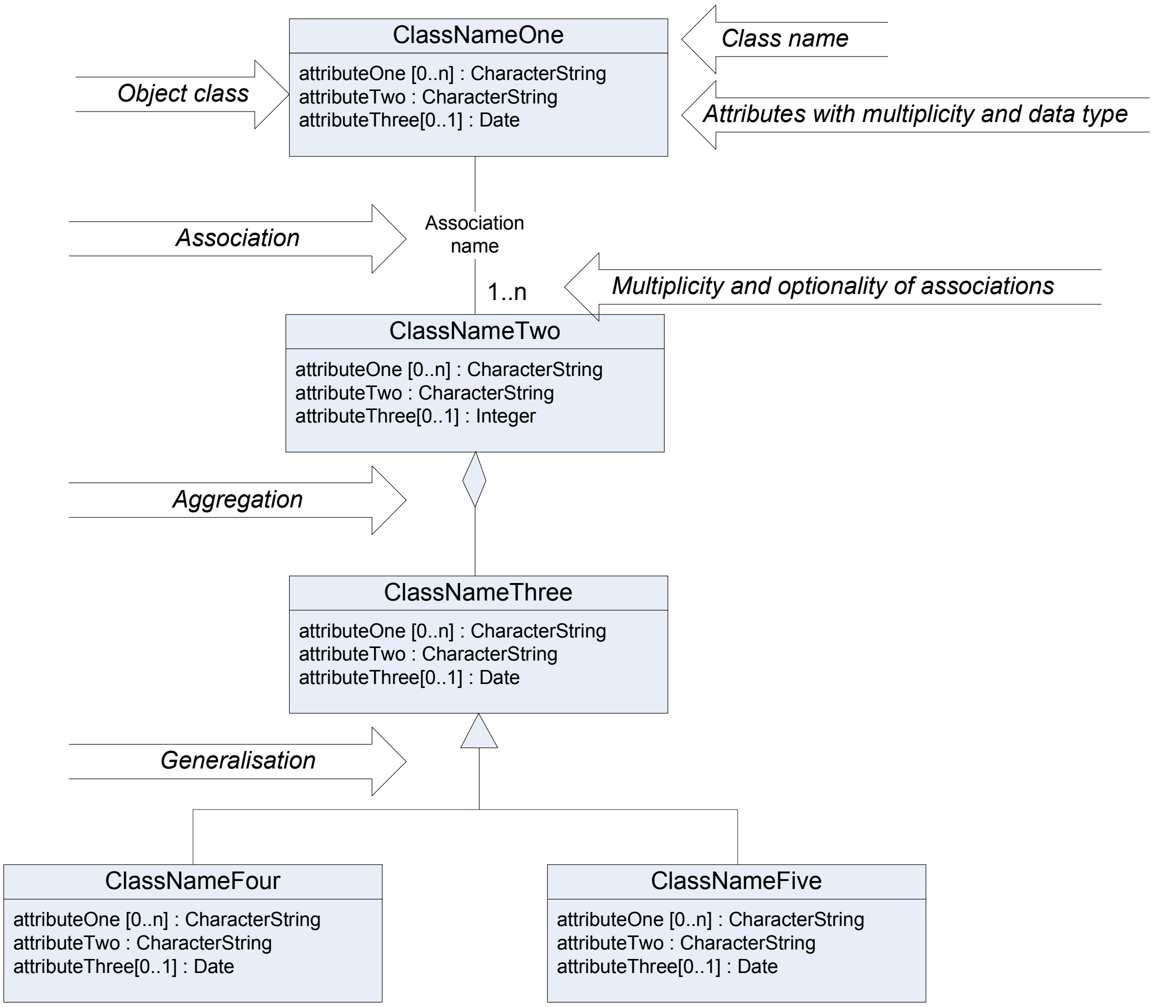
ClassNameOne
attributeOne [0..n] : CharacterString
attributeTwo : CharacterString
attributeThree[0..1] : Date
Class name
Object class
Attributes with multiplicity and data type
Association
Association name
Multiplicity and optionality of associations
1..n
ClassNameTwo
attributeOne [0..n] : CharacterString
attributeTwo : CharacterString
attributeThree[0..1] : Integer
Aggregation
ClassNameThree
attributeOne [0..n] : CharacterString
attributeTwo : CharacterString
attributeThree[0..1] : Date
Generalisation
ClassNameFour
attributeOne [0..n] : CharacterString
attributeTwo : CharacterString
attributeThree[0..1] : Date
ClassNameFive
attributeOne [0..n] : CharacterString
attributeTwo : CharacterString
attributeThree[0..1] : Date

Figure 3. UML diagramming

References

BS 7666-0 Spatial datasets for geographical referencing – Part 0: General Model

BS 7666-1 Spatial datasets for geographical referencing – Part 1: Specification for a street gazetteer

BS 7666-2 Spatial datasets for geographical referencing – Part 2: Specification for a land and property gazetteer

BS 7666-5 Spatial datasets for geographical referencing – Part 5: Specification for a gazetteer of delivery points

ISO 639-2 Codes for the representation of languages – Part 2: Alpha-3 Code

ISO 8601 Data elements and interchange formats 0 Information interchange – Representation of dates and times

ISO 19112 Geographic information – Spatial referencing by geographic identifiers

ISO 19113 Geographic information – Quality principles

ISO 19114 Geographic information – Quality evaluation procedures

Section 2. How to create a gazetteer of a new type of geographic object

This Section of the Guidelines covers how to create a gazetteer of a new class of geographic object. Part 0 of BS 7666 defines a generic gazetteer model that can be used for any class of geographic object, and this section describes how to implement this. It describes the planning and initiation of such an implementation and general principles of gazetteer construction. Finally the process of creating a new part of BS 7666 is outlined. A Glossary of Terms, a list of abbreviations and references, and an explanation of the UML data modelling convention used in the Standard are given in Section 1 of the Guidelines.

1. Introduction

BS7666 has become synonymous with streets and land and property units. However, since its inception, it has been the long-term intention to extend the Standard with further parts specifying gazetteers of other types of geographic object. This has been done with the new Part 5, for delivery points. The purpose of Part 0 is to define a generic gazetteer model that can be used for any type of geographic object. It can be applied to any type of geographic referencing object, i.e. ones that are used for referencing and locating other business information.

A gazetteer can be produced for any type of geographic object. What will differ between them is the form of spatial reference used.

Examples of geographic objects that are within scope:

- Country
- Administrative area
- Census output area
- Town
- Locality
- Watercourse (rivers, lakes, etc)
- Harbour
- Industrial site
- Military site
- Hospital
- Public buildings
- National park
- Environmental protection area
- Contaminated land
- Planning zone
- Development land
- Estate land

Note that these do not have to be national but may be restricted to a specified territory and may be restricted by ownership, i.e. either public or private.

2. Planning and initiation

2.1 Definition of scope

Before starting any gazetteer implementation, it is necessary to define the scope of the gazetteer. This should describe the types of object to be included in the gazetteer, with the rules for inclusion or exclusion. In defining the gazetteer scope, it is necessary to take into account why the gazetteer is required, what purpose it will serve, and the type of information that will be linked to geographic objects contained in it.

Example of gazetteer scope
Localities in Great Britain, defined as named neighbourhoods, suburbs, districts, villages, estate or settlements (but not towns) that are used to reference several properties, whose name is recognised by the local authority for purposes of addressing.

2.2 Data specification

BS 7666 defines the general structure of the gazetteer. It does not define the details of content. As part of any implementation, it is necessary to specify in detail the data to be included. This should include the following:

- **Description of the details of the implementation of the Standard**, defining the data structures to be used (use of additional attributes, field lengths, domains, classification systems and codes etc);
- **Rules for inclusion of instances of the geographic object** and their identification;
- **Identification of the source of the data**: a process needs to be established to recognise, identify and label instances of the geographic object. This is likely to be from existing data sources, for example records of properties recorded for a particular purpose. It should include new instances being created, where a well-defined business process needs to be established, including how the new instances are referenced, for example a process of naming or address creation;
- **Identification of the attributes of each class of geographic object**: this may include ones additional to the requirements of the Standard.

2.3 Object classifications

Objects in the gazetteer may not all be the same type. Two levels of classification scheme are allowed. The primary classification scheme should be used for high-level, external purposes, to roughly divide the objects.

Example classification of localities:
Localities to be captured in a gazetteer could be classified by:

1. Rural villages or settlements
2. Suburbs of towns
3. Industrial or trading estates
4. Other

A secondary classification scheme can be used for more detailed classifications, such as internal or application-specific classifications. Note that the secondary classification scheme need not be a refinement of the primary classification scheme, and may be independent of it.

2.4 Data maintenance

A gazetteer is not a static dataset, but a continually changing description of a set of real-world objects. Consequently, it is essential that a maintenance regime is established. There are three main stages in the life-cycle of an object record in the gazetteer, creation, change and closure. Different procedures are required for each.

- **Creation**: a business process needs to be devised to identify new instances of the geographic object in the real-world, and to collect the necessary data about them. This will involve some level of interaction with the life-cycle of the object, for example notification by some regulatory body.
- **Change**: change to a gazetteer entry can occur for many reasons. They essentially fall into two categories, those representing real-world change and those due to correction of current data or insertion of missing data..
- **Closure**: a business process needs to be devised to identify when instances of the object cease to exist in their recorded form. This may involve some level of interaction with a regulatory body. The gazetteer record for this instance is then amended to change its state, and to input a value for the end date. Historic records should not be deleted, as they may be still be of interest, but may be archived.

2.5 Data quality

Quality levels and the processes required to control and assure that these levels are maintained need to be established at the outset of gazetteer creation. How this can be done is described in more detail in Section 3.

3. General principles of gazetteer construction

3.1 Mandatory, optional, conditional and additional attributes

In the Standard, attributes are identified as being mandatory, optional, or in some cases conditional.

Mandatory attributes should always be provided. **Optional** attributes should be provided, except where they are not applicable, e.g. for additional languages in a non-English Gazetteer, or where what is referred to does not exist or is not required, e.g. some additional fields in an address. In some cases, an attribute is recorded in the Standard as being **conditional**, and a condition is stated when it should be provided. In this case the attribute is mandatory when the condition is satisfied.

In certain instances it may not be possible to determine the value of an attribute. Null values for mandatory attributes should be avoided since it is ambiguous whether the value has been omitted in error or is unknown.

Additional attributes may be provided in an individual implementation. These should be specified in the data specification.

3.2 Spatial references

A spatial reference is a description of a real-world place which can then be used to reference other information. A coordinate reference is a particular type of spatial reference, but here we are primarily concerned with non-coordinate references, based upon names of real-world places. The best example of this is an address. A general address structure is given in Annex C of Part 0, and is used in Parts 1 and 2. Depending on the type of object included in the gazetteer, and the territory of use, variations on this may be required.

Examples of spatial referencing system

- Geographic address: object name or number, street name, locality name, town name, administrative area name (see section on **Gazetteers and Addressing**);
- Postal address;
- Postcode;
- County name;
- Country name or code.

Where a new spatial referencing system is required for a gazetteer, it should be defined as specified in **4.2.4** of Part 0 of the Standard.

3.3 Geographic extents

Geographic extent is a detailed description of the "footprint" of the geographic object, recorded either as a collection of one or more geographic objects or as one or more boundary polygons. Where the extent is described by a collection of smaller geographic objects, the identifiers of these are recorded. An example of where this might be used, is Regions defined as collections of counties and other local authority areas, where the extent is described by the identifiers of these local authorities. Where the extent is described by boundary polygons, these may be recorded by either a set of coordinates, or by polygon identifiers such as TOIDs (identifiers for topographic objects in Ordnance Survey's MasterMap® product). Multiple polygons may be used for non-contiguous areas, and "cut-outs" may be used to exclude inner polygons, for example for some current counties. These are illustrated in Figure 1.

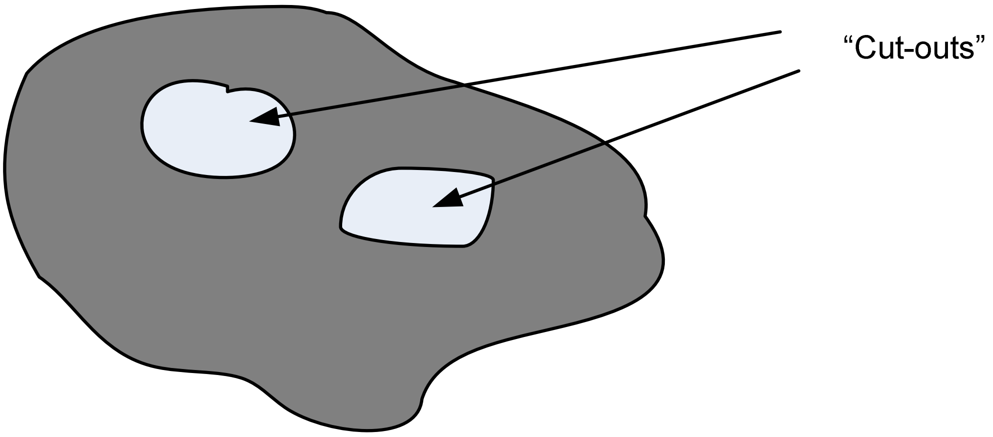

Figure 1: "Cut-outs" within a polygon

3.4 Dates

Dates should be recorded at an appropriate level of resolution. Normally this will be a day, but where this is not known, it may be only a month (e.g. ‘2006-08’) or a year (e.g. ‘1900’). The dates should be recorded consistently either in the basic format (YYYYMMDD) or extended format (YYYY-MM-DD, YYYY-MM or YYYY), where YYYY is the year, MM the month and DD the day. The two formats should not be mixed, and for each implementation of the Standard it will have to be decided which format is to be used.

Care needs to be taken to distinguish between actual dates, when something happened either in the real world or to a source of information, and capture and update dates, when changes are made to the data.

Actual dates:

- **start date:** the date at when the geographic object came into existence;
- **end date:** the date when the geographic object ceased to exist;
- **current state date:** the date when the geographic object came into its current state;
- **current date:** the date at which the gazetteer is considered to be current.

Capture and update dates:

- **entry date:** the date when the geographic object record was entered into the gazetteer;
- **update date:** the date when the geographic object record was last updated;
- **metadata date:** the date when the metadata was last updated.

Where dates are not known exactly, a notional date at which the date criterion was known to be correct should be used. This applies in particular to start date. Where update date is not known, or the record has not been updated, the update date should be same as the start date.

3.5 Links to other objects and gazetteers

In common with the other Parts of the Standard, Part 0 specifies a facility for linking to other objects and gazetteers. How this is implemented will depend on the nature of the object and the corresponding related objects and also the nature of the linkage. The linkage is always geographic in nature but the objects do not have to be coterminous or exactly coincident spatially. The relationship may involve time as well as space.

In all cases, the link will take the form of the identifier or identifiers of the related object or objects in the other dataset or gazetteer. The nature of the relationship and the dataset to which the data is cross-referenced should be identified in the metadata (see section on **metadata**). This relationship need not be one-to-one (i.e. an object in one dataset may relate to more than one object in another dataset. The correspondence need not be exact as described above).

Examples of possible types of relationship:

simple cross-referencing: objects such as land parcels correspond to those in another dataset. The cross-reference is then between the respective identifiers of the same instance of the object in the gazetteer and the other dataset.

Temporal: an object now occupies the same location as that occupied historically by another object, for example, a plot of land which has now been built on. The cross-reference is between the identifier of the historic object and that of the object which has replaced it (wholly or in part).

Complex cross-referencing: the objects within scope of the gazetteer are similar but not identical to those recorded in another dataset, for example the objects in a land and property gazetteer and the topographic features on a digital map. The cross-reference is then between the identifier of the gazetteer object and those of the instances of the topographic features that most closely represent the object.

Parent-child: one object (the child) cannot exist without a corresponding higher level object (the parent). The child is often formed through subdivision of the parent object. The parent and child may be the same class of object, for example Primary Addressable Objects and Secondary Addressable Objects such as a flat within a larger property or different classes of object, for example streets and elementary street units. Explicit provision is made in the standard for parent-child relationships.

Other geographic objects: for example objects in the same gazetteer whose geographic footprint overlaps, for example buildings on different vertical levels.

These relationships are illustrated in Figure 2.

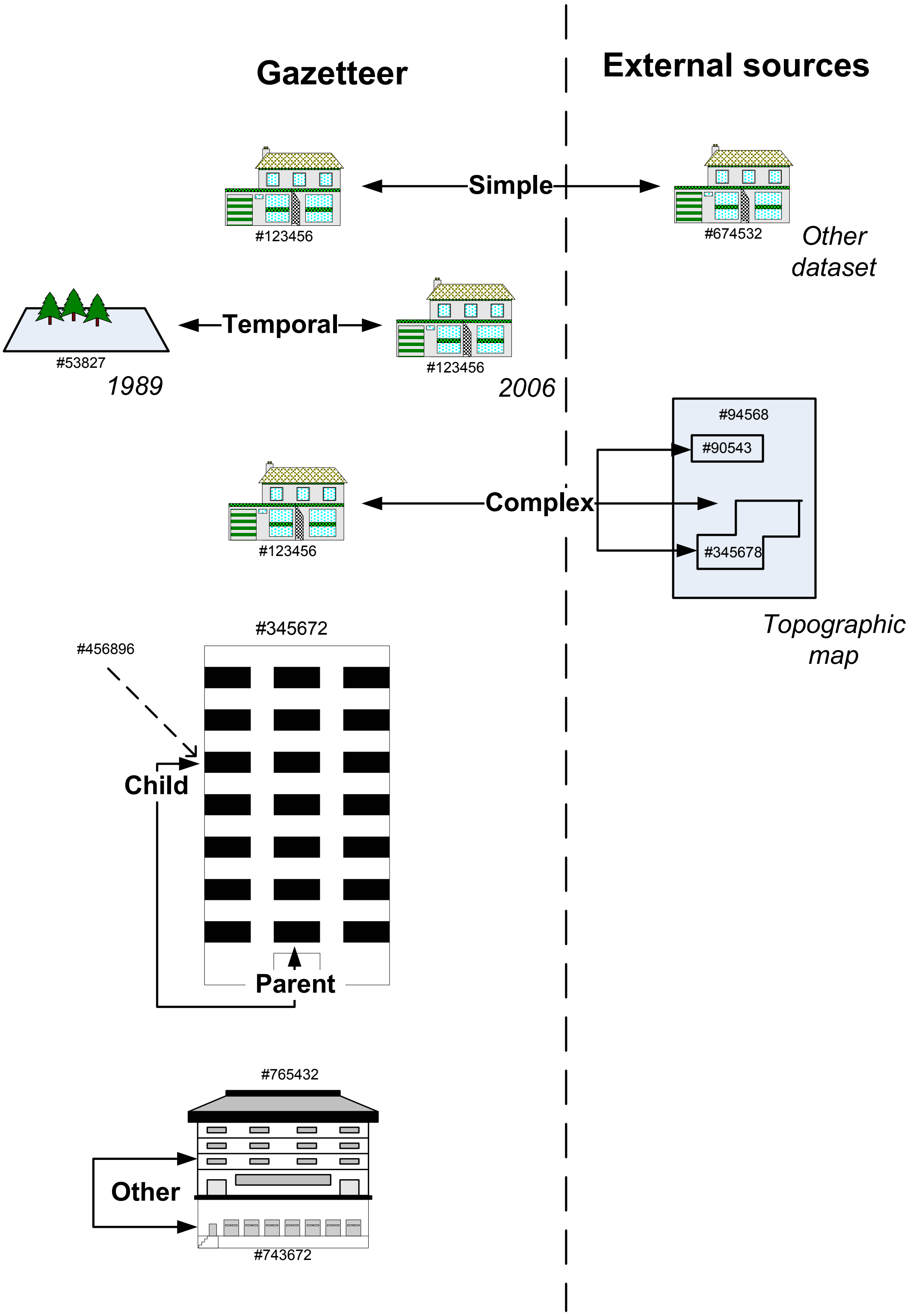

Figure 2: Possible types of relationship which can be recorded in a gazetteer

3.6 Metadata

The Standard mandates the provision of metadata for all gazetteers. The main purpose of the metadata is to describe the dataset. It not only provides basic information for operational purposes, but can be fed into a metadata service such as *gigateway*[2] for general data resource enquiries.

Clause **4.2** of Part 0 of the Standard specifies various metadata items:

- **Name**: the formal name of the gazetteer;
- **Scope**: a description of the types of objects to be included in the gazetteer e.g. trunk roads;
- **Territory of use**: the area covered by the gazetteer, e.g. 'England and Wales', 'Great Britain', 'Scotland', 'Northern Ireland';
- **Language**: any non-English languages used in the gazetteer, for example Welsh. Note that this refers to the language used for making the entry not the words themselves, as for example an English language gazetteer may contain words in other languages. If only the English language is used, then this need not be stated;
- **Character set**: any non-English character sets used, e.g. Gaelic. If only the standard character set is used (as for English), then this need not be stated;
- **Gazetteer owner**: the organisation with overall responsibility for the gazetteer. Note that this may not be the same as the gazetteer custodian;
- **Gazetteer custodian**: the organisation or officer responsible for the compilation and maintenance of the gazetteer;
- **Coordinate system**: the system used to define coordinates of points in the gazetteer. This is likely to be the National Grid of Great Britain for Great Britain, and ITM (Irish Transverse Mercator) for Northern Ireland. Note that these are land based systems, and gazetteers that cover wider areas are likely to use some form of latitude and longitude (e.g. ETRF 89);
- **Coordinate axis units**: the unit of measurement for the coordinates, for example metres;
- **Metadata date**: the date of the last update to the metadata;
- **Spatial referencing system**: this is likely to be specific to the type of object included, but in general will be some sort of address or set of references to higher level units such as countries;
- **Primary and secondary classification schemes**: the classification schemes used;
- **State coding scheme**: any scheme used to describe the operational state of objects in the gazetteer, for example the stage reached in its lifecycle;

[2] The UK GI metadata service - see gigateway.org.uk

- **Current date**: the date at which the gazetteer is considered to be current. Note that this is different from the date of last update of the data;
- **External cross-referencing schemes**: any external cross-referencing schemes used.

Some of these metadata elements are mandatory, and some are optional. Optional items should be provided whenever they are applicable and known. Multiple values may be given for language, character set, spatial referencing system and external cross-referencing scheme, where appropriate.

3.7 Data quality

The Standard specifies a requirement for a data quality report. This is an assessment of the data in terms of the following:

- Lineage
- Currency
- Positional accuracy
- Attribute accuracy
- Completeness
- Logical consistency

These are discussed in more detail in Section 3 of these Guidelines.

4. How to create a new part of BS 7666

4.1 The need for further Parts

BS 7666 defines gazetteers for various types of geographic objects. There are specific Parts covering gazetteers of streets (Part 1), land and property (Part 2) and delivery points (Part 5). Further Parts could be created for gazetteers of other types of geographic objects if required.

Part 0 can be used to define a gazetteer of a different type of geographic object. As with Part 5, it might be considered desirable to make a specific additional Part of the standard to cover this. This is likely to be done when:

- There is a clear business case;
- There is a widespread requirement in a specific user community;
- There will be implementation in a range of organisations;
- There are specific spatial referencing mechanisms that need to be clearly defined;
- There is a need to collect data on a local basis for creation of a national dataset.

4.2 The standards creation process

BS 7666 comes under the auspices of BSI Committee IST/36 *Geographic information.* This committee proposes and provides technical approval of British standards. It is sponsored by the Association for Geographic Information (AGI).

The stages in the creation of a standard are as follows:

- Business case approval: a formal proposal is made to BSI, to ensure that the proposed standard meets their requirements, and on approval leads to formal set-up of a Standard development project;
- Obtain resources: because of the formal nature of a standard, professional drafting is required, and funding for this may be obtained through DTI, which can take some time;
- Set-up Steering Group: this will contain representatives of the major stakeholders and will be responsible for the overall conduct of the project;
- A series of drafts are produced for review by a Working Group, culminating in a Draft for Public Comment (DPC);
- Public consultation: the DPC is published by BSI, and formal comments are invited from the public;
- Review of comments: a Review Group goes through all the comments received, produces a response to them, and makes any required changes to the draft standard;
- BSI conformance: BSI Editorial Department ensure that the standard conforms to the rules for standards, and arranges the presentation of the document;
- Technical approval: IST/36 reviews and approves the standard for its technical content;
- Publication: this is done by BSI who determine the price of the publication, and provide some marketing.

After five years, all standards are subject to a formal review. This may recommend one of the following:

- To ratify the standard for a further five years;
- To issue technical amendments;
- To create a new edition of the standard;
- To withdraw the standard.

Section 3. Quality assessment and reporting

This section of the Guidelines covers quality assessment and reporting. It introduces the principles and concepts of data quality in the context of gazetteer creation, maintenance and utilisation and provides general guidance on how to test and report on gazetteer quality in conformance with BS 7666: 2006. This includes a description of some basic data quality measures and quality evaluation procedures, some quality test methods, the outline of a data quality report and an overall process for controlling quality of a gazetteer. It is at a generic level, and will be expanded in later sections in the context of particular types of gazetteer. A Glossary of Terms, a list of abbreviations and references, and an explanation of the UML data modelling convention used in the Standard are provided with Section 1 of the Guidelines.

1. Introduction

All users of a gazetteer need to be aware of the degree to which it is current, complete, accurate and consistent. They also need to understand the gazetteer's history and how and why it was derived so that they can judge the suitability for their purposes. In other words, they need to know something of the quality of the gazetteer data.

All parts of BS 7666: 2006 include the requirement that the quality of data in a gazetteer must be tested and reported. Each gazetteer must have associated with it a data quality report containing details of the results of tests covering aspects of the quality of the data.

The purpose of this section of the Guidelines is to introduce the principles and concepts of data quality in the context of gazetteer creation, maintenance and utilisation and to provide general guidance on how to test and report on gazetteer quality in conformance with BS 7666: 2006.

Since neither this Section nor any other section of the Guidelines relates to any specific implementation of the Standard, acceptable levels of quality cannot be recommended. However, this Section gives general guidance to users of the Standard on designing tests and arriving at acceptable levels of quality for their own implementations. Further details relevant to particular parts of the Standard are to be found in those sections of the Guidelines dealing with these parts.

This Section of the Guidelines does not cover the managing of quality of national gazetteers compiled from local gazetteers. This topic is dealt with in Section 4 of the Guidelines.

Just as for manufacturing, software development or service provision, quality should not be an "add-on" to gazetteer production. It should not be a cursory check at the end of the process. To create and maintain a gazetteer which is fit for the purpose, quality needs to be built into the processes of data capture, compilation and update. Given the right tools, procedures and above all, people trained in the processes, this need not be exacting. Other downstream processes are likely to rely on the quality of the gazetteer. They will incur additional costs if the gazetteer does not meet acceptable levels of quality or, in common parlance, is not "fit for purpose".

2. The management of quality

2.1 What is quality?

Gazetteers are a type of geographic dataset; they record information about locations on the earth's surface. To achieve an acceptable level of quality in any geographic dataset requires careful control of the processes involved in their creation and maintenance and adequate testing to ensure that acceptable levels of quality have been achieved.

This sub-section sets out some basic concepts in the management of quality. These concepts and their application to gazetteer data are explored in more detail in the next sub-section.

First we need to establish what we mean by quality. Quality can be defined as:

- 'Fitness for purpose';
- 'Performance against specification';
- 'Totality of characteristics of a product that bear on its ability to satisfy stated and implied needs'[3].

In the case of gazetteer data conforming to BS 7666, this could be redefined as:

- Fitness for use in the locating of information to a specific geographic place;
- Degree of conformance to BS 7666 and the data specification for a particular implementation of a gazetteer;
- Suitability of a gazetteer for particular purposes.

Quality is frequently described by terms such as "correct", "comprehensive", "accurate", "consistent" and "reliable". These are all subjective terms - "accurate" for one user is "inaccurate" for another. Likewise, saying that all geographic data has to be of the "highest quality" to ensure that all needs are satisfied is largely meaningless because, once again, it is a relative term. If by "highest quality" is meant that all data must be completely accurate, correct and consistent then this is both unrealistic and unachievable. It is unrealistic because it implies that all records need to be checked against the real world as it existed at a particular point in time. It is unachievable because the effort required in attempting to reach "100% quality" is out of all proportion to the value added.

2.2 Achieving acceptable quality in geographic data

To achieve an acceptable quality in geographic data we need to have:

- quality designed into the data creation and maintenance process – in other words – provide quality assurance;
- clear quality goals relating to specific aspects of quality (e.g. accuracy, completeness, consistency) which are based on user needs;
- quality control of the creation and maintenance processes to ensure that errors are minimised;
- adequate testing at the end of the process to provide assurance that quality levels are being achieved;

[3] International Standard - ISO 8402: 1994 Quality Management and Quality Assurance – Vocabulary.

- clear roles and responsibilities for quality management.

None of the above is possible without a comprehensive data specification. Without this, it is impossible to construct measures and tests and provide an overall assessment of the quality of the data because there is no detailed indication of what you are testing against.

Further, we need to be able to tell the users about the quality of the data. Thus a system of quality reporting is needed.

2.3 Establishing quality goals

To establish quality goals it is necessary to analyse quality into a number of quality aspects that are applicable to the data. These can include, for example, positional accuracy, attribute accuracy and completeness. We need to arrive at acceptable quality levels for each of these, in doing this they need to be SMART i.e.

- **S**pecific – they should absolutely clear about the aspects of quality or the quality elements for which levels are being set;
- **M**easurable – an acceptable quality level has to be measurable or else it cannot be evaluated effectively;
- **A**chievable – setting levels which are unachievable has no purpose since all data will fail;
- **R**ealistic – this means that there needs to be a compromise between what can feasibly be achieved by the data producer, what can feasibly be tested and what is deemed acceptable by the user;
- **T**imely – in the sense of being expedient and practical e.g. to conduct tests to measure the quality level.

2.4 Quality control of the creation and maintenance process

Quality is not an "add-on" at the end of a process. To best meet specified quality levels, quality needs to be built into the data creation and maintenance processes; it is not just a question of testing at the end of the production process.

"**Quality control**" refers to the way in which quality checks are carried out during the production process and items failing quality checks are managed.

During data creation and maintenance, there may be a number of checks carried out most notably at data entry. Depending how the data is compiled it may be possible to build in the validation of certain quality components such that all mandatory entries are completed, all values fall within specified domains and certain basic consistency checks are made. If any data does not pass one of these checks then there have to be procedures for identifying, quarantining and dealing with the failures.

2.5 Quality testing

To ensure that the data reaches acceptable quality levels and to be able to report data quality, tests need to be devised. These can take a number of forms:

- tests of some aspect of all records in the dataset;
- tests of a sample of records, usually chosen at random, of the whole dataset;
- comparisons with the original source of the data such as the real world;

- comparisons with other sources of information regarded as being true such as maps or other geographic data;
- automatic tests using software – these tend to be used to establish the degree of internal consistency within a dataset;
- manual tests using inspection – these tend to be used where original sources or other information needs to be compared with the dataset.

Should the dataset fail to meet the specified quality levels then it should not be released but should be corrected and retested.

2.6 Reporting quality

A standard approach to reporting quality needs to be adopted, preferably using some type of *pro forma*. This should include:

- information about the dataset,
- how it was produced or the lineage,
- how up to date it is or its currency,
- what aspects of quality were tested,
- details of the tests (complete dataset or sample, automatic or manual),
- the source information used for the tests (original source or secondary) and
- the results.

Where a dataset is being maintained and changes are frequent, it will not be practical to update the quality report after every change and a decision will have to be made about what triggers another report, elapsed time (e.g. monthly, annually) or amount of change (e.g. change to 5% of records).

2.7 Quality roles and responsibilities

Meeting acceptable quality levels in the end comes down to people, people to design and manage the whole process, people to capture and maintain the data and people to test the data. For the management of quality to be effective, there need to be clear roles and responsibilities.

- The owner of the data should ensure that:

  (i) a comprehensive data specification is in place,

  (ii) clear quality goals are set which are going to meet user needs, and

  (iii) those managing the data production process have the skills and are adequately resourced.

- The manager of the production process needs to:

  (i) design the process such that there are adequate quality controls and a final test process to be able to deliver data meeting the agreed quality levels,

  (ii) ensure that those capturing and maintaining the data have the right tools, instruction and training to do the job,

  (iii) have in place mechanisms for acting on user feedback, and

  (iv) always be seeking quality improvement and ways of raising quality levels.

- Those producing and testing the data need to be quality aware and understand the quality levels required.

The users also have a key role; they after all will be looking at individual records and will inevitably spot errors and inconsistencies. This needs to be exploited and users encouraged to feedback any perceived errors. However, unless these users see the feedback acted upon, they will cease to provide it and confidence in the dataset will be eroded.

3. Managing the quality of gazetteers

3.1 Data quality and gazetteers conforming to BS 7666

Having introduced some basic concepts of managing the quality of geographic data we now go on to examine how these concepts can be applied to gazetteers conforming to BS 7666: 2006.

This sub-section concludes with a more general discussion about managing quality including quality roles, quality control and quality improvement.

Aspects of quality relating to particular types of gazetteers are given in the Guidelines specific to parts 0, 1, 2 and 5 of BS 7666: 2006.

3.2 What is special about managing the quality of gazetteers?

Gazetteers present particular challenges when trying to manage their quality. Notable amongst these are:

- the volumes of records that need to be compiled in a gazetteer – even for a local gazetteer this can involve many thousands of records and for a national gazetteer, many millions;
- the diverse sources of information – apart from the real world, there are address lists, existing records, maps, all created for different purposes at different times and all having different qualities;
- the difficulty of ensuring and enforcing consistency in recording when geographic objects, even within one class, can be so varied – for example the diversity of land and property objects;
- the lack of consistency in the naming, numbering and describing of objects in the real world – this can result in a lack of consistency in records;
- maintaining consistency when creating and maintaining gazetteers where a number of people may be involved who are spread across several work areas;
- the dynamic nature of many types of gazetteers – the geographic objects recorded in a gazetteer will be changing rapidly – for example business premises;
- the user requirement for a high level of currency, completeness and accuracy in a gazetteer such that they can link their information to the locations recorded in a gazetteer with assurance.

(These problems can be compounded if local gazetteers are amalgamated into national gazetteers. For a discussion of these problems see Section 4 of the Guidelines.)

3.3 Errors frequently found in gazetteers

Because of the nature of gazetteers, certain types of errors tend to be prevalent. Many of these errors can be overcome by a more rigorous approach to managing quality. These errors include:

- excessive division of geographical objects such as streets (including paths) adding little or no value for the user and complicating the maintenance of the gazetteer;
- entering the same object in a gazetteer multiple times leading to duplicate records – for example the same street having several “unique” street reference numbers (USRNs);
- objects being entered which are out of the scope of the gazetteer - examples being railway lines or canals masquerading as streets, or lampposts and bus shelters masquerading as land and property units;
- objects referencing other objects which do not exist e.g. Basic Land and Property Units (BLPUs) referencing streets which are not included in the relevant street gazetteer;
- inconsistencies of identifiers, for example names of geographic objects and spatial units;
- missing mandatory attributes, such as spatial references;
- streets with an incomplete descriptive identifier e.g. missing locality, town or administrative area such that the street cannot be located and identified uniquely within the territory of use;
- inconsistencies and misapplication of parent-child relationships, for example primary and secondary addressable objects.

3.4 Requirements in BS 7666

BS 7666-0: 2006, Clause 6 “Data quality” states that the quality of data in a gazetteer must be tested and reported. A data quality report is required recording a standard set of data quality measures associated with each gazetteer, and containing details of:

- any tests carried out – including details of the test methods;
- the date of each test;
- the name of the tester;
- details of any source material or other information used in the testing.

Particular sections and their contents are specified in respect of the following quality aspects:

- lineage;
- currency;
- positional accuracy;
- attribute accuracy;
- completeness;
- logical consistency.

Aspects of data quality to be included in a quality report

The quality of gazetteers can be described in terms of a number of distinct components or elements referred to as "aspects" in BS 7666: 2006. These can be divided into two types:

Descriptive – these provide information and background to the production and purpose of the dataset so that the user may judge the general suitability of the dataset to their particular application;

Quantitative – these describe numerically, as a percentage or as a pass/fail, how well a dataset meets the criteria set out in its product specification and enable the user to determine whether the quality levels achieved meet their quality requirements.

In the case of BS 7666: 2006, the descriptive aspects are:

Lineage – this describes the history of the dataset, the sources used, the maintenance applied and the methods used in the derivation of the data and changes made since its inception.

Currency – this gives an indication of how up-to-date the gazetteer is by providing a date for which the gazetteer is current.

In the case of BS 7666, the quantitative quality aspects are:

Positional accuracy – closeness of the stated positions to those accepted as true. Although not specified in BS 7666, positional accuracy can be either:

- **Absolute accuracy** - closeness of reported coordinate values to values accepted as being true (e.g. coordinate values on an Ordnance Survey large-scale map);
- **Relative accuracy** – closeness of the relative positions of features in a dataset to their respective relative positions accepted as being true (e.g. position relative to the extent of a street or a Basic Land and Property Unit).

Attribute accuracy – correctness of the values, including dates, entered for each attribute. The Standard distinguishes between discrete attributes i.e those carrying discrete values such as a classification code or an address and continuous attributes i.e. those carrying any value within a specified range such as a date. Detailed quality aspects relating to discrete attributes are:

- **Classification correctness** – the correctness in the application of a particular class to an item;
- **Non-quantitative correctness** – the correctness of an entry which does not contain a quantitative or class value in relation to that believed to be true (e.g. a descriptive identifier);

And that relating to continuous attributes is:

- **Quantitative accuracy** – the accuracy of a value expressing a quantity - this includes dates.

Completeness – degree to which the data is complete in respect of the stated gazetteer scope and the currency. Although not termed as such in BS 7666, this can include errors of **omission** (items missing) or **commission** (items duplicated).

Logical consistency – degree of consistency to rules for the recording and encoding of data items within the dataset. These can include **conceptual consistency** (e.g. is object within the scope of the gazetteer?), **consistency of association** (e.g. does this record correctly reference another record?), **domain consistency** (attribute values fall within a permitted range of values), **format consistency** (data is of the specified format including data type) and **temporal consistency** (dates are correctly ordered e.g. update date not earlier than entry date).

These requirements are echoed in the other parts of the standard relating to streets, land and property and delivery points where similar or identical wording is used for the data quality clauses.

The verification of conformity of a gazetteer to BS 7666 includes a check that it has a quality report meeting the requirements set out the Data Quality clause.

It should be noted that lineage and currency are descriptive quality aspects, there are no tests associated with these aspects.

3.5 What needs to be implemented

The Standard only outlines a requirement for the reporting of data quality, the detail needs to be specified in any implementation of a gazetteer.

The Standard sets out data quality measures for some quality aspects although in no case is the test method specified. These need to be developed as part of an implementation.

The data quality aspects which have specified measures are as follows:

- Positional accuracy - accuracy of coordinates in the gazetteer in terms of distance on the ground;
- Attribute accuracy:

  - Discrete attributes - percentage found correct;

  - Continuous attributes - mean error;
- Completeness:

  - Entries included in the gazetteer in accordance with its stated scope expressed as a percentage of those which should have been present at the date when the gazetteer was current;

  - Duplicate entries included in the gazetteer expressed as a percentage of all entries as of the date when the gazetteer was current.

The above should not be taken to mean that other measures cannot be used in addition. No data quality measures are specified for logical consistency for example, these have to be developed.

Again the Standard does not state whether the whole population of records has to be tested or a sample. If random[4] methods are used, the method of generation of the sample needs to be reported however.

There is no requirement in BS 7666: 2006 to report on the quality of the metadata although normal practice would be to include tests of the metadata content as part of an overall test strategy for a gazetteer.

[4] Methods of random sampling and the reporting of results based on sampling are out of scope of these Guidelines.

Data Quality Evaluation

There are a number of terms used in relation to assessing the quality of datasets. Aspects of quality have already been described.

Data quality scope – this refers to the degree of applicability of a data quality aspect to the data. The aspect may relate to all attributes and associations in a complete gazetteer but more likely it will relate to:

- only some classes of geographic objects (e.g. metalled public roads), certain attributes (e.g. descriptive identifier) or associations (e.g. streets and their subdivision into elementary street units);
- a particular geographic extent – only part of the territory of use of a gazetteer;
- a particular temporal extent – only those records collected within a certain time frame (e.g. between 1st June 2004 and 21st July 2005).

Data quality measure – this is what is measured about each of the data quality aspects. For example if the data quality aspect is completeness then we may choose to measure the number of missing records in a dataset as a percentage of those entered. A data quality measure may have several tests associated with it.

Test methods or data quality evaluation methods – these are how the data is tested to provide a particular data quality measure. These test methods may be applied to all the records or data items within the data quality scope or a sample. They may compare the data with the real world or some other dataset.

Test results or data quality results – these are the results from applying the test methods. The results may be expressed as percentages, as a mean error or root mean square error, or as simple pass/fail depending on the data quality measure and whether the test is against an acceptable quality level.

Thus in any implementation of a gazetteer conforming to BS 7666: 2006, the following needs to be specified:

- the aspects of quality to be assessed (as a minimum this has to be those given in the Standard) ;
- the data quality scope or scopes relating to a quantitative data quality aspect;
- the data quality measures (these have to include those specified in the Standard);
- the test methods including the sources of information for any tests;
- acceptable quality levels;
- the exact form of the quality report;
- how and when quality reports are to be issued e.g. at every release of the gazetteer, every 6 months, after so many changes to gazetteer entries.

Details are given in later parts of these Guidelines dealing with specific Parts of the Standard.

3.6 Data quality measures and users' priorities

In addition to the data quality measures mandated in the Standard, it is necessary to prioritise what other data quality measures are to be included and then specify their test methods.

It is not going to be feasible to measure all possible quality aspects for all possible attributes and associations. Apart from basic computing requirements for domain and format consistency and some degree of referential integrity, priorities have to be set for assessing the quality of the data content

In setting priorities, it is essential to consider the user. Beyond currency, completeness and accuracy of position, it is suggested that for the user, the accuracy and consistency of certain attributes is going to be a priority, notably the:

- spatial references (e.g. Land and Property Identifier, descriptive identifier of a street) so that they can recognise and confirm that the location described in the gazetteer is correct.

Other priorities may be:

- Update date – to indicate the likely currency of the record.

Of lower priority are likely to be (for example):

- Administrator;
- Secondary classification.

4. Test methods

4.1 Introduction

Having arrived at a number of data quality measures for the gazetteer, then specific test methods need to be devised to arrive at a data quality result.

In deciding on suitable tests, there are a number of basic issues that need to be resolved:

- What is going to be the basis for evaluating quality, are external sources available (other datasets or the real world) or can we only test against the data itself?
- Can some or all of the methods be automated?
- Are tests going to be based on a full inspection of the data or only a sample?

Test methods[5].

Tests or data quality evaluation methods can be divided into two main types, **direct** and **indirect.**

With **direct methods** data quality is determined by making a comparison of the data with various reference information. This information can be either:

- **internal** - contained within the data itself. For example, tests of domain and format consistency need only the data itself;
- **external** – using whatever is available other than the dataset itself. For example tests of positional accuracy or content correctness require information from other datasets or the real world.

Indirect methods infer or estimate data quality using external information on the data such as data sources and reports or knowledge of the data production process.

To measure effectively all elements of data quality in a gazetteer, direct methods using both internal and external information sources are likely to be required although indirect methods may be necessary when compiling national gazetteers (indirect in the sense of having to rely on the quality reports supplied with local gazetteers – see Section 4).

For some types of measures, the tests may be susceptible of **automation**, this is typically so in the case of direct methods using internal sources. Measures of logical consistency frequently fall into this category where data types or domains are being tested. In other cases **visual inspection** will be needed.

Full inspection involves the testing of every data record within scope. Typically, full inspection is relevant to small populations of records or automated methods.

Larger datasets such as gazetteers are likely to have to use **sampling** where manual inspection is required, such that sufficient items are tested to give a meaningful data quality result.

These Guidelines do not include specific guidance on sampling methods; this is a major topic in its own right. The reader is referred to the relevant ISO Standard on quality evaluation[6].

Examples of test methods relevant to the data quality measures discussed above are given in Table 1. Under test method, the use of random sampling is proposed.

[5] What follows is based in part on International Standard, ISO 19114:2003 Geographic information – Quality evaluation procedures.

[6] International Standard, ISO 19114:2003 Geographic information – Quality evaluation procedures.

Data quality aspect	Detailed data quality aspect	Data quality scope	Data quality measure	Test method
Positional accuracy	Absolute accuracy	Positions of all geographic objects recorded in the gazetteer	Mean error in metres of recorded coordinates of representative points against those believed to be true	1. Take each point from a random sample of all geographic object records. 2. Determine by inspection of a large-scale OS map if each point is correctly positioned according to the data specification. 3. Where point is incorrect, measure distance from where point is believed to be and convert to distance on the ground. 4. Take all test results and compute the mean error. (Mean error = the mean value of all the errors without regard to sign.)
Attribute accuracy	Non-quantitative correctness (of discrete attribute)	Spatial references of all geographical objects recorded in the gazetteer	Number of location records with incorrect spatial references as a percentage of all location records	1. Take a random sample of all geographic object records. 2. Determine by inspection of a large-scale Ordnance Survey map and any other external sources if each spatial reference conforms to the data specification. 3. Take all test results and compute a percentage of those incorrect.
	Quantitative accuracy (of continuous attribute)	Start dates of all location records in the gazetteer where not notional dates (i.e where the true start date is not known)	Mean error of the start dates of all location records in the gazetteer where this is known	1. Select those geographic object records that have start dates that are not notional dates. 2. Determine as far as possible from external sources the accuracy of the start dates and determine the difference between that and the recorded date. 3. Take all test results and compute a mean error for the overall result. (Mean error = the mean value of all the errors without regard to sign.)
Completeness	Omission	All geographic objects within scope of the gazetteer and the territory of use existing on the date the gazetteer was current.	Geographic objects present in gazetteer as a percentage of those present in the real world (or believed to be present) at the time the data was current	1. Take a random sample of areas within the territory of use. 2. Determine by inspection of a large-scale Ordnance Survey map, aerial photography of an appropriate date and any other external sources (or, if necessary, by visits on the ground) whether each geographic object existing in the sample areas at the date when the gazetteer was current had been captured. 3. Take all test results and compute those geographic objects recorded in the gazetteer as being present in the sample areas as a percentage of those actually believed to have been present when the gazetteer was current.

	Commission	All geographic objects within scope of the gazetteer and the territory of use existing on the date the gazetteer was current.	Duplicate entries in the gazetteer for the same geographic object as a percentage of those geographic objects present in the real world at the time the data was current	1. As above but determine by inspection if the same geographic object recorded more than once (albeit with different identifier). 2. Take all test results and compute those geographic objects duplicated in the gazetteer in the sample areas as a percentage of those actually believed to have been present when the gazetteer was current.
Logical consistency	Conceptual consistency	Records of all geographic objects recorded in the gazetteer.	Geographic objects outside of gazetteer scope (i.e. not of the types as specified in the scope) as percentage of all geographic objects recorded in the gazetteer	1. Take each record from a random sample of all records. 2. Determine by inspection of each record (and, if necessary, any other external sources available) if the geographic object is within the geographic scope. 3. Take all test results and compute the number of geographic objects recorded which are outside the gazetteer scope as a percentage of all geographic objects recorded in the sample.
	Association consistency	All attributes that cross-refer to other attributes within this or associated gazetteers[7] (cross-references to external entities are out of scope)	Pass if all cross-references refer to valid items existing in the gazetteer or associated gazetteers	1. Pass all records through a software validation which checks that all cross-references exist in the gazetteer or an associated gazetteer. 2. Fail the gazetteer if any failures.
	Format consistency	All attributes in all geographic object records in the gazetteer	Pass if all formats (maximum occurrences, data types) consistent else fail	1. Pass all records through a software validation which checks that all attributes are consistent with the format rules set out in the data specification. 2. Fail the gazetteer if any failures.
	Domain consistency	All attributes in all geographic object records in the gazetteer	Pass if all values fall within valid domains else fail	1. Pass all records through a software validation which checks that all attributes are consistent with the domains set out in the data specification. 2. Fail the gazetteer if any failures.
	Temporal consistency	All attributes with data type of date in the gazetteer.	Records with incorrect date ordering (e.g. start date earlier than entry date: entry date earlier than update date) as a percentage of all records	1. Pass all records through a software validation which checks that all dates are correctly ordered. 2. Compute the number of records that fail as a percentage of all records.

Table 1: Examples of data quality aspects, scopes, measures and test methods applicable to gazetteers

[7] For example land and property gazetteer and its associated street gazetteer.

4.2 Development of acceptable quality levels for gazetteers

Testing the quality of a gazetteer and reporting the results is a valuable exercise in its own right. However, it is far more valuable if the gazetteer quality is compared to pre-determined quality levels such that the dataset is not released unless these levels or quality goals are reached. As stated previously, acceptable quality levels or AQLs need to be SMART i.e. **S**pecific, **M**easurable, **A**chievable; **R**ealistic and **T**imely. Organisations may aspire to "100%" quality but it can never be achieved, better to state a realistic and achievable level and ensure that it is met.

BS 7666: 2006 does not specify AQLs, these need to be established as part of a gazetteer implementation. In setting AQLs, the user must always be borne in mind and their priorities made paramount. As discussed above, the users' priorities are likely to be completeness, and the accuracy and consistency in the recording of spatial references.

Acceptable quality levels (AQLs)

Acceptable quality levels (AQLs) are threshold values applied to the results of testing data quality to determine whether the data meets criteria determined from a data specification or user requirements.

AQLs can be based on various types of values such as Boolean (true or false), numeric or percentage depending on the types of measures adopted.

AQLs can be applied to the results from tests for a specific component of quality such as 'absolute accuracy' and be specific to one attribute such as 'type of property' to determine whether that attribute meets the specified criteria. These are called **simple** AQLs.

Alternatively, AQLs can be applied to the aggregated results from a number of tests to determine whether a gazetteer meets the specified criteria[8]. These are called **aggregated** AQLs. For example, aggregated AQLs could be:

- 100% pass/fail – all attributes in a dataset must reach or exceed the AQL for each attribute;
- Weighted pass/fail – all data quality results are weighted and scored according their perceived significance. Those not achieving a threshold score are deemed to have failed;
- Subset pass/fail – only those data quality results considered important e.g. all mandatory elements, must pass to achieve an overall pass.

These Guidelines do not propose any specific AQLs since these are implementation dependent. More specific guidance is given in those Sections dealing with Parts 1, 2 and 5.

[8] Derived in part from International Standard, ISO 19114: 2003, Geographic information – Quality evaluation procedures, Annex J

5. Reporting gazetteer quality

5.1 Requirements

The Standard does not mandate the form of the quality report although it mandates that certain details are included, these are set out in **3** above.

The general form of a quality report that would be suitable is given in **5.3** below. This assumes that tests were completed on a single version of the gazetteer current at a stated date. The report does not include any item for metadata, if this is tested this should appended to the report.

5.2 When to report gazetteer quality

Gazetteer data will not be static; the real world which is represented in the data is continually changing. The approach to quality reporting that is adopted will depend on how and when the changes are made available to the user, two possible scenarios are:

1. the changes are made available as soon as the location records are created or updated – albeit with some degree of quality control (see below);
2. the changes are made to a production gazetteer with formal periodic releases of a user version.

In the first scenario it will not be feasible to repeat tests on the whole gazetteer and update the quality report after every change or even after a number of changes. In effect, some sort of **benchmarking** will be needed, that is taking a copy of the gazetteer periodically e.g. once a week, once a month, after 100 changes depending on how much change there is and the size of the gazetteer. The copy of the gazetteer is then tested and reported upon as if it was static. If the testing shows that the acceptable quality levels are no longer being achieved, then action will have to be taken to correct the location records (see below).

In the case of the second scenario a more controlled approach will be possible and tests can be run on the whole gazetteer and assurance gained that acceptable quality levels are being maintained before release of the next user version of the gazetteer. This approach is preferable but may not be feasible in a particular business context where updates are required daily.

5.3 Data quality report applicable to gazetteers conforming to BS 7666

(notes on each entry are given in italics)

Covering page

This page should appear on the front of all reports

Name of gazetteer:	*As stated in the metadata e.g. "Land and Property Gazetteer of the City of Winchester".*
Scope of the gazetteer:	*As stated in the metadata – e.g. "public streets and streets used for creating addresses of residential and commercial property".*
Date of quality report:	*Date the report compiled, not the date of testing, this is indicated against individual tests.*
Report compiled by:	*Name and position of the compiler of the report e.g. "J Smith, Street Gazetteer Custodian, Borsetshire County Council".*
Owner of gazetteer:	*As stated in the metadata e.g. "Borsetshire County Council".*
Testing organisation	*Organisation which actually conducted the tests, it may be the organisation owning the gazetteer, a department within the organisation or some separate organisation. Details should include name and address and contact for the organisation.*
Additional information	*Add any other information relevant to the gazetteer quality e.g.:* • *whether the report relates to the initial creation of the gazetteer or is an update of a previous report following changes to the gazetteer;* • *for a national gazetteer compiled from local gazetteers, whether this is a summary of local testing or the result of testing at the national level.*

Summary of results for each quality aspect

Additional summary information may be added, that shown is for minimum conformance to BS 7666. If this is a national gazetteer compiled from local gazetteers and the results are indirect (e.g. aggregated from local quality reports), then this should be made clear in the summaries below.

Lineage of gazetteer:	*The history of the dataset, the sources used, the maintenance applied and the methods used in the derivation of the data and changes made since its inception. Provide sufficient information such that a potential user can form an initial judgement of the applicability and value of the gazetteer.*
Currency at time of testing:	*This is date at which the gazetteer was considered current at the time of testing e.g. 9th June 2006.*
Positional accuracy	*Summary of result(s) of tests of positional accuracy of the coordinates in the gazetteer in terms of distance on the ground. State whether based on sample or the whole dataset.*
Attribute accuracy	*Summary of the results of tests carried out on:* *1. the accuracy of the discrete attributes in the gazetteer expressed as the percentage found correct;* *2. the accuracy of continuous attributes in the gazetteer expressed as a mean error.* *List the summary results by attribute tested. State whether based on sample or the whole dataset.*
Completeness	*Summary of the result(s) of tests to verify:* *1. that all entries have been included in the gazetteer in accordance with its stated scope - express results as a percentage present;* *2. that there are no duplicate entries - express as a percentage of duplicates.* *State whether based on sample or the whole dataset.*
Logical consistency	*Summary of the results of tests to verify that entries in the gazetteer have been recorded in a consistent manner.* *List the summary results by type of consistency tested e.g. format consistency, domain consistency. Express results as a pass or fail or percentage failures as appropriate.* *State whether based on sample or the whole dataset.*

Details of tests
Each page should contain information about one test. Entries for detailed quality aspects, data quality scopes and data quality measures may need to be repeated where multiple scopes, measures or tests are employed.

Quality aspect	*As a minimum there should be reports on tests of positional accuracy, attribute accuracy, completeness and logical consistency*
Detailed quality aspect	*These are subdivisions of the major quality aspects, for example:* • *for positional accuracy - absolute positional accuracy;* • *for attribute accuracy - classification correctness, non-quantitative correctness;* • *for completeness - commission and omission;* • *for logical consistency - conceptual consistency, association (cross-referencing) consistency, temporal consistency, domain consistency and format consistency.*
Data quality scope	*Degree of applicability of a data quality aspect to the data. Indicate whether this is all attributes and relationships in the gazetteer, only some data items (e.g. specific attributes, certain associations), only some types of geographic objects (e.g. streets but not elementary street units), a particular geographic area which is only part of the territory of use or only a certain time frame.*
Data quality measure	*Details of what is measured about the data quality aspect (e.g. if the detailed data quality aspect is completeness – omission then the measure might be records omitted as percentage of those present in the gazetteer).*
Test method	*A description or a reference to the method used to apply the data quality measure to the data quality scope*
Internal/external sources	*Describe the information sources used i.e. internal - contained within the data itself or external – outside sources of information. If the latter, then the sources should be described e.g. OS Land-Line, aerial photography, Postcode Address File.*
Automatic/manual	*Indicate whether automatic e.g. software, manual inspection or both were used for test.*
Full inspection/sampling	*State whether there was a full inspection of the data or a sample, in which case the method of generating the sample should be described or referenced.*
Test result	*Give in units as stated in the data quality measure.*
Date of test	*This is the date the test was actually run.*
Name of testers	*List all those involved starting with the lead tester.*
AQL	*Acceptable quality level if available or applicable.*
Pass/fail	*Where applicable, indicate whether the acceptable quality level was passed.*

6. Overall processes for controlling quality

6.1 Introduction

As was stated in the introduction to this Section, quality is not an “add-on”. To maintain and improve quality levels, quality needs to be built into the gazetteer creation and maintenance processes. It is not just a question of testing at the end of the production process and then relying on the user to report errors.

The overall planning of production processes to ensure that the product meets the required quality levels is often referred to as **quality assurance**. **Quality control**[9] refers to the way in which quality checks are carried out during the production process and how items failing quality checks are managed.

During data creation, maintenance and release or supply to users there may be a number of checks carried out:

- at data entry – depending how the data is compiled it may be possible to build in validation of certain quality components such that all mandatory entries are completed, all values fall within specified domains and certain basic consistency checks are made;
- prior to release to the user – more comprehensive checks are made using manual and automatic checks.

If any data does not pass one of these checks then there have to be procedures for identifying, quarantining and dealing with the failures.

The production of gazetteers may be in the context of a broader “quality management system” in operation throughout the whole organisation. This goes further than quality assurance and embraces all those activities needed to deliver quality i.e. planning, operations, evaluations and staff training. There is a strong emphasis on prevention rather than correction and continuous quality improvement.

6.2 Gazetteer creation and maintenance

A simple flow model of gazetteer creation and maintenance is presented in Figure **1**1 which shows the quality related processes.

It is assumed that:

1. the gazetteer has already been created and that there is a working or production copy of the gazetteer as well as the released copy available to users;
2. the gazetteer is formally released in versions to users.

The flow is much idealised and relates to the general case and not to a particular class of locations such as streets. Processes and the sequence in which they are performed may vary widely between specific implementations.

(The processes involved in building national gazetteers from local gazetteers are not included. These are discussed in Section 4 of the Guidelines.)

[9] The terms “quality assurance” and “quality control” are frequently interchanged leading to an erosion of meaning. The terms are used in the way defined here throughout the remainder of these Guidelines

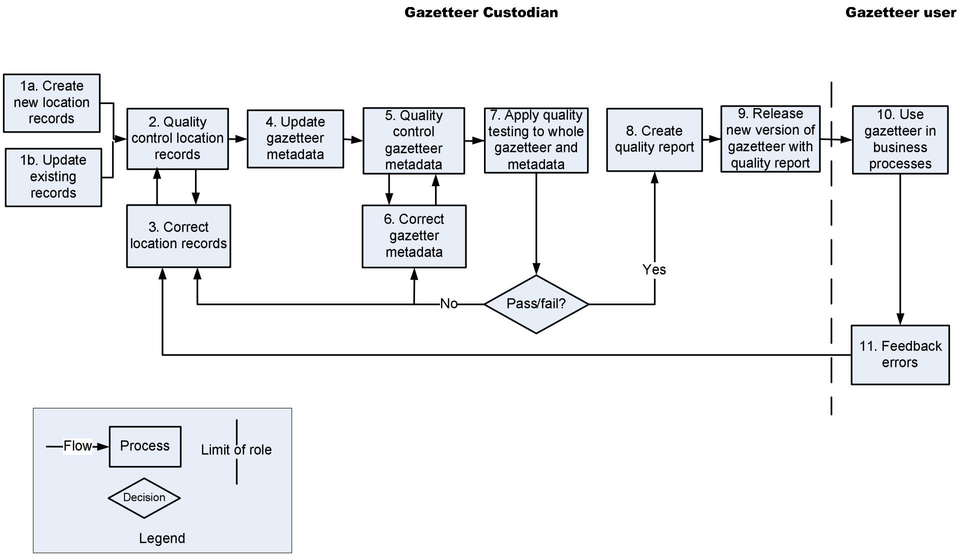
Gazetteer Custodian
Gazetteer user
1a. Create new location records
1b. Update existing records
2. Quality control location records
3. Correct location records
4. Update gazetteer metadata
5. Quality control gazetteer metadata
6. Correct gazetteer metadata
7. Apply quality testing to whole gazetteer and metadata
8. Create quality report
9. Release new version of gazetteer with quality report
10. Use gazetteer in business processes
11. Feedback errors
Pass/fail?
No
Yes
Flow
Process
Decision
Limit of role
Legend

Figure 1: Simplified flow diagram showing gazetteer creation and maintenance

The terms and processes used in the diagram are described in more detail below. The model serves to provide an overview and a context.

1. Location records are created (1a) or updated (1b) in the production or working copy of the gazetteer as the first stage.
2. The location records are quality controlled by manual inspection or validation at entry or both. If by inspection, this may be by random sampling within batches. If any batch fails then the whole batch is quarantined and checked.
3. Location records failing quality control are corrected and subjected to further quality control until the batch passes.
4. The gazetteer metadata is updated if required.
5. The gazetteer metadata is also subject to quality control.
6. If the metadata fails quality control it is corrected and rechecked until it passes.
7. The whole gazetteer is subjected to a comprehensive testing quite independent of the production process using different staff. If the tests show that any of the quality aspects do not meet acceptable quality levels then the location records or the metadata or both need to be corrected and resubmitted for final testing.
8. If the acceptable quality levels are achieved then a quality report is compiled.
9. The new version of the gazetteer is released with the quality report.
10. Users make use of the new version of the gazetteer in their business processes.
11. Users' feedback any errors which are corrected as part of the process of creating the next version of the gazetteer.

The key points to derive from this flow diagram are the need:

- to make quality integral to the gazetteer creation and maintenance process;
- for an overall approach to quality assurance with clear points at which there are quality controls;
- for documented quality evaluation procedures and agreed AQLs;
- for good feedback loops from user to gazetteer custodian.

Two quality roles are shown on the flow diagram, gazetteer custodian and gazetteer user. These are described in the following paragraph.

6.3 Quality roles for gazetteers

The two primary roles for managing gazetteer quality are:

- the **gazetteer owner** – the organisation that has overall responsibility for the gazetteer;
- the **gazetteer custodian** - the organisation or individual that is responsible for the compilation and maintenance of the gazetteer.

(These roles may be performed at a higher level where local gazetteers are amalgamated into regional or national gazetteers. These national roles are discussed in more detail in Section 4 of the Guidelines.)

A third role, often overlooked, is:

- the **gazetteer user.**

These roles, in general terms, are described below.

The gazetteer owner is responsible for ensuring that:

- a comprehensive data specification for the gazetteer implementation conforming to BS 7666 is in place;
- clear quality goals in the form of acceptable quality levels are set which are feasible to maintain and will meet user needs;
- there is an overall quality assurance which is properly planned and implemented and involves quality control and final and independent testing;
- those managing the gazetteer production process have the skills and are adequately resourced.

The gazetteer custodian is responsible for assuring the quality of the gazetteer. Key responsibilities are likely to include:

- ensuring that detailed capture and maintenance instructions and guidance are in place;
- understanding the quality requirements for the gazetteer;
- having clear procedures for dealing with errors and user feedback;
- establishing or agreeing AQLs that meet user requirements (or at least can be practically achieved) with the gazetteer owner;
- providing quality assurance through flowline design with adequate quality control built-in;

- ensuring that adequate procedures are in place for both gazetteer creation and maintenance;
- ensuring that suitable tools are available for gazetteer capture and maintenance which conform to BS 7666 and the data specification for the implementation;
- having quality control and final testing procedures backed by suitable testing tools;
- providing for the identification, quarantining and correction of gazetteer data failing quality evaluation;
- providing for adequate training such that staff have an understanding of the purpose of the gazetteer and are familiar with the capture and test tools;
- having a quality reporting procedure which ensures that quality reports conforming to BS 7666 are available when the gazetteer is released and during maintenance;
- having change control procedures;
- creating a culture of quality improvement where feedback is encouraged and lessons are learned and applied.

The gazetteer users are the ultimate beneficiary. Without the users there is no point in creating a gazetteer or building a service around it. They do not have to be passive players in this; they can be tremendous source of free (and unsolicited) comment and informal quality control. They will experience inconsistencies and notice errors. By submitting comments they can contribute to the improvement process but it has to be seen that the gazetteer custodian is pro-active in this area or else the users will cease to submit feedback and become more disgruntled with the gazetteer.

6.4 Maintaining and improving data quality

Users will be looking for continuing improvements in data and this means that there will be a drive for further quality improvement. To achieve this there need to be mechanisms for:

- feeding back and acting on errors found in data;
- feeding back on improvements to processes by operators of those processes;
- learning and applying lessons from the use of current processes;
- managing change whether to the data specification or the AQLs.

To use that old adage “prevention is better than cure”, it is always going to be more cost-effective to create data that meets AQLs than to have to correct it every time. This is likely to be achieved through a combination of better procedures, tools and staff training. No-one knows better than the staff doing the job, they deserve to be heard and responded to where they have ideas for improving the process.

Section 4. How to create a 'national' gazetteer

This Section of the Guidelines deals with how to create a 'national' gazetteer. It describes how to amalgamate 'local' gazetteers into a 'national' one. It identifies quality issues between the two types of gazetteer and in particular issues relating to how a national gazetteer should be maintained. A Glossary of Terms, a list of abbreviations and references, and an explanation of the UML data modelling convention used in the Standard are given in Section 1 of the Guidelines.

1. Introduction

1.1 Local and national gazetteers

Gazetteers, especially street gazetteers and land and property gazetteers, are often described as being 'local' or 'national'. These terms were chosen to correspond to levels of government involved in the creation and usage of such gazetteers. Land and property gazetteers for example, are being created in local authorities and merged to form the National Land and Property Gazetteer (NLPG). However, here we will use the terms 'local' to mean of restricted coverage, and 'national' to mean of more widespread coverage encompassing several 'local' gazetteers.

1.2 Creation of a national gazetteer by amalgamation of local gazetteers

There is a further assumption that 'national' gazetteers are created by amalgamation of local gazetteers. This need not necessarily be the case. For example, a national gazetteer of streets could be created directly from an existing source such as Ordnance Survey data. In this section we will concentrate on how to amalgamate local gazetteers into 'national' ones, although many of the issues identified are also relevant to creation of national gazetteers directly.

1.3 Purpose of a national gazetteer

The purpose of a national gazetteer and a local gazetteer could well be different. For example, objects of local interest may be of no significance nationally. Indeed the local priorities might vary in different areas, resulting in inconsistencies nationally. Similarly, the degree of detail on which locations (particularly streets and BLPUs) are recorded could be quite different, resulting to issues of granularity. Typically, local gazetteers require all used locations to be included, whilst national gazetteers only require locations that are significant on a wider basis.

1.4 Issues of national consistency

When local gazetteers are created separately by different independent organisations, it is inevitable that there will be variations in content, consistency and quality, due to different capabilities, resources and commitment amongst those creating and maintaining the data. Many of these issues are to do with people, management, organisation and copyright, which are outside the scope of these Guidelines.

A nationally agreed and implemented specification is essential; the existence of the Standard alone cannot ensure consistency. Whereas local specifications may be extended according to local requirements there must be conformance with the national specification. The nature of national data specifications is described below.

An exchange format for the transfer of data to the national level will be highly desirable. This combined with a software validation process to make basic checks on data formats and domains will considerably facilitate the creation of a national gazetteer.

It will also be a big help to have tools for capture, validation and output of data which conform to the national specification and can be shown to do so. Ideally, this should be through a national certification process whereby tools are taken by a neutral party and subjected to an agreed set of controlled tests.

Those creating the local gazetteers need to understand the national requirements. Thus, training and adequate instructional material and guidance produced at the national level are likely to be needed.

It is important to ensure user feedback from any usage at the national level and a means of rapid correction at the local level.

2. How to amalgamate 'local' gazetteers into a 'national' one

2.1 Definition of national scope

The scope for a national gazetteer will be different from that for a local gazetteer. These differences could include extended territory of use, more restricted rules for inclusion, and reduced currency. These should be identified in the scope statement for the national gazetteer.

2.2 National data specification

As for any gazetteer implementation, a detailed data specification is required, conforming to the standard. At the national level, this should include the following:

- Description of the details of the implementation, defining the data structures to be used;
- Rules for inclusion of instances of the location and their identification;
- Implementation-specific classification schemes to be used, for example standard lists of codes used in the description and naming of geographic objects e.g. DfT road classifications, ONS local authority code lists;
- National referencing schemes to be used, establishing consistent supporting geographies where possible for a national system of spatial units, for example:
    - standard lists of localities and towns to be used,
    - inclusion of prefixes on identifiers, to make them nationally unique,
    - inclusion of higher-level fields in an address;
- Identification of the attributes of each class of location to be provided;
- Specification of cross-references required;
- Target levels for gazetteer currency with frequency of submission of updates.

2.3 Capture and maintenance rules

As for any gazetteer implementation, practical guidance is required for such things as:

- Procedures for initial capture of data;
- National agreement and mechanisms for the assignment of unique identifiers;
- Procedures for maintenance of data, including frequency of update;
- Statement of what objects may be held at a local level, but is not required at the national level (which may be identified by means of ‘flags’ in the implementation). These excessive items should be filtered out before submission to the national dataset;
- Establishment of nationally recognised “standards”, against which data can be tested (e.g. Ordnance Survey large-scale mapping for coordinates);
- The format for transferring the data to the national level;
- National quality standards, specifying Acceptable Quality Levels, which cover not only logical consistency in terms of format and domain, but also other aspects set out in the Standard, including associations;
- Details of quality control and quality checks to be carried out locally, including methods, measures, reports and Acceptable Quality Levels.

2.4 Data differences at the national level

Some objects may be different at the national level compared with their occurrence in local gazetteer. Some may need to be amalgamated into true ‘national’ objects, for example trunk roads, motorways, rivers and BLPUs that cross local boundaries.

Descriptive identifiers may provide a unique spatial reference within the local territory of use but they may not be unique nationally. For example a locality or town name may be unique in a particular county but not nationally, this should be covered in the national data specification but checks will still need to be made.

2.5 Quality control and quality assurance

The quality of any national gazetteer compiled from local gazetteer is going to be heavily dependent on quality management at the local level. Acceptable quality levels for local gazetteer submissions will have to be determined and agreed together with the method of testing to establish that levels are being met.

On initial submission, each local gazetteer should have a quality report. This will need to be updated from time to time as agreed nationally. A protocol will have to be in place for dealing with data that does not meet nationally agreed quality levels.

Even with these processes established, it will be necessary to assure the quality of the data submitted to achieve as much consistency as practical. This will have some limitations. Checks on the ground or even of local data sources are unlikely to be feasible. Periodic audits of both local quality management processes and data could be considered if the value of the national gazetteer justifies it.

It will not be possible to replicate all the tests carried out at the local level or to establish that all the acceptable quality levels have been met. A more limited set of acceptable quality levels will need to be established which tie-in to those at the local level.

Most tests at the national level are going to have to rely on the gazetteer data itself, typically tests of logical consistency. For content correctness, nationally available datasets such as PostcodeAddress File (PAF) and Ordnance Survey MasterMap or Land-Line can be used but these have currency and scope limitations. It is suggested that the following at least should be checked at the national level:

- **Logical consistency**
    - **Format:** is the data in the correct format, i.e. are all fields of the specified data type (this can be checked by software processes);
    - **Domain:** do the attribute values fall within expected ranges (this can be checked by software processes), these might include:
        - Reference numbers in pre-defined ranges,
        - Codes from the national pre-defined lists,
    - **Temporal consistency:** date not before a certain date, and not in the future (this can be checked by software processes),
    - **Association consistency:** are cross references valid (it should be possible to check these by software processes). For example:
        - Streets referenced by BLPUs,
        - Primary addressable objects referenced by secondary addressable objects.
- **Completeness –** this is likely to require manual checking of a sample and comparison with other national datasets:
    - **Commission** : are there entries for objects or locations that are out-of-scope such as lamp posts in a land and property gazetteer, or duplicated (particularly due to overlaps of data scopes between adjoining local gazetteers);
    - **Omission**
        - are the total number of entries in the gazetteer at the level expected for the current date. This will difficult to determine precisely but alternative sources should indicate the approximate number expected;
        - are all mandatory attributes present (this can be checked by software processes);
- **Attribute accuracy** - for many attributes, it is possible to define a set of valid values (some software checking or software tools may assist with this). These might include:
    - names, by checking single occurrences of names used in spatial references;

- **Positional accuracy -** are the individual coordinate references for the locations within the territory of use of the local gazetteer (software could be used for this if the polygon defining the territory of use of the local gazetteer is known).
- **Granularity:** are the objects at the right level of granularity, for example are streets excessively subdivided (this will require manual checking).

However, a great deal can be gained from an initial overall assessment. A visual examination of the raw data in the local gazetteer will give confidence, or otherwise, in the quality of the data.

Further details of tests are described in the section on Quality Assessment and Reporting. At the conclusion of all testing of a local gazetteer, a test report should be sent back to the data originator. This should identify the tests performed, who performed them and when, and the results. For further details of test reports, see the section on Quality Assessment and Reporting. Any errors found, either specific or general, should be identified, requesting that they be investigated and corrected. Any datasets failing to reach the Acceptable Quality Level should be rejected.

Where tests can be performed by software, these could be carried out by the supplier of the data as a pre-condition for submission of the data. This requires nationally available software tools for data validation.

To conform to BS 7666 it will be necessary to issue a quality report for the national gazetteer. This needs to summarise the quality that can be expected at the national level. The accuracy of this report will depend upon:

- the rigour of the quality controls and quality assurance in the production and maintenance of the local gazetteers;
- the achievement of nationally established acceptable quality levels locally;
- the accuracy of the quality reports accompanying the local gazetteer submissions;
- the checks applied to the local gazetteers on receipt at the national hub.

National users will expect national consistency in the quality of the data. Reporting average values for quality measures which mask large variations at local level will not endear users to the national gazetteer.

3. Maintenance

3.1 Introduction

Data maintenance is important in all gazetteers. For a national gazetteer, the issues are more than just processing updates from the local gazetteers. There are additional issues of consistency, currency and frequency of update. There is likely to be great variability in the maintenance regimes for the different local gazetteers. Whilst it is relatively straightforward to obtain updates on a regular basis, it is likely that the currency of these updates (when the real-world change occurred as opposed to when the local gazetteer was updated) will vary considerably. This will lead to variations in quality across the national gazetteer.

3.2 Processing change from local gazetteers

Changes resulting from changes in the real world will be of three basic types:

- **Changes to existing instances:** the changes will need to be checked, including validity of the identifiers;
- **New instances:** as well as checking the individual entry, its relationship to any other location will need to be checked, to ensure referential integrity;
- **Closure of instances:** any references to the location in other gazetteer entries will need to be checked, to ensure that no orphan records are created, for example a street should not be deleted while there are BLPUs that reference it. Deleted entries should be archived to provide a historic record.

There will also be changes resulting from the correction of errors. These can be any of the above.

A record of history should be maintained through periodic archiving of versions of the gazetteer.

3.3 Change at the national level

Some objects in the national gazetteer may not occur as such in a local gazetteer, or may occur in part in different gazetteers. These may include motorways, high-speed rail links etc. Separate process will need to be set up to deal with changes to these. Notice of some of these changes may come from sources other than local gazetteers.

3.4 Quality control

As well as checking change data on receipt, additional testing of the data against that already held will be required. This should particularly check for issues of consistency, for example duplicate entries being introduced, deletions of entries not held nationally, conflicts with other entries and changes being reversed. The methods outlined earlier will be applicable here. Most queries will have to be referred back to the data source for resolution.

From time to time, the national quality report will have to be re-issued, as indicated above. This will rely heavily on local quality management and reporting. See the Section on Quality Assessment and Reporting for more details.