First World War Thesaurus - Appendix 1

1. Data Standards and Terminology Control

1.1 What are Data Standards?

Data standards are sets of rules and conventions which encourage the recording of information in a consistent and retrievable way. They are a statement of what data should be recorded, how it should be recorded and the ways in which it can be supported within a system in order to retain its full meaning. The development and application of a data standard is vital to ensure that users can access and retrieve data not only within specific systems but also across a range of systems operating within an organisation. It is possible through the use of agreed standards and terminology control to ensure the consistency of information held within a data set.

1.2 Terminology Control Mechanisms

When dealing with data of any kind, it is essential that the information contained within a database can be readily retrieved and understood by anyone. By standardising the way in which information is entered into the database it is easier to search the records and retrieve the data required. In a database, each field will relate to a specific concept and therefore any term entered into a field should fall within its definition; if a field relates to survey the user should only expect survey types to be entered/displayed within that field. Also, it is necessary to introduce some form of terminology control to ensure that data entered by one person can be retrieved by another. The simplest way to ensure that the information is consistent is to use a wordlist. This is simply an alphabetical list of accepted terms used to control the information recorded in a specific field within a database. However, a wordlist does not allow the user to create relationships between the terms.

Below is a wordlist containing various types of First World War sites, each of which could be used to index the records:

EXPLOSIVES FACTORY CENOTAPH PRISONER OF WAR CAMP MILITARY CAMP
BORING MILL NAVAL BATTLEFIELD OBSERVATION POST COACH WORKS
COMMEMORATIVE STONE CHARGE HOUSE BLAST WALL RIFLE RANGE

If a user is only interested in retrieving the records for military sites from the database, then searches on at least five separate site types are required to retrieve all the information and even then the user needs to be aware of any abbreviations or punctuation used in the entries when making the search. This is only a short list and already retrieval has become a lengthy, time-consuming process. By using this thesaurus structure, expanding abbreviations and removing punctuation the number of searches required is automatically reduced.

1.3 What is a Thesaurus?

A thesaurus is used to standardise terminology and help the user to choose terms to enter into a field. However, unlike a wordlist, a thesaurus:

a) allows terms, related by a similar subject, to be grouped together into hierarchies and cross-referenced to other groups of terms which may be relevant to the subject.
b) provides the user with a single preferred term to use where there is a choice of terms with the same or similar meaning, for example: **SOUND MIRROR** use for Sound Dish.
c) through the use of hierarchies, allows terms to be selected at a general or specific level, depending on the level of indexing required.
d) is a dynamic tool, which can be developed by the addition, amendment and deletion of terms, relationships or hierarchies as dictated by individual needs.

Where sets of data relate to the same (or similar) subjects, a thesaurus can form the standard for information held across a number of data sets managed by different organisations. This enables a user to interrogate any number of databases which use the thesaurus, safe in the knowledge that the information they require will be presented using a terminology they are familiar with.

1.4 Thesaurus versus Wordlist

Consideration should be given as to whether it is necessary to produce a thesaurus as its construction is more resource intensive than a wordlist and therefore it may be simpler to retain a wordlist (if one exists). However, a thesaurus has a number of advantages when dealing with large data sets, namely:

a) it increases retrieval and eliminates redundant data through the use of the hierarchical structure and associative relationships.
b) it enables a system to be used by several indexers and searchers within an organisation, whilst maintaining a consistent level of indexing.
c) it enables indexing and searching to be carried out at either a general or specific level depending on the detail of information available/required.

2. Structure

The structure of this thesaurus is based on guidelines given in the British Standard BS5723: 1978 *Guidelines for the establishment and development of monolingual thesauri* and the third edition of *Thesaurus Construction* (Aitchison, Gilchrist and Bawden, 1997). It deviates from these standards in that:

a) it uses the singular form rather than the plural. This decision was based on the fact that most heritage recording bodies use the singular form in their databases.

and

b) it groups terms by class rather than the broadest noun term (Top Term). It was felt it would be useful to group terms under CLASS schemes thereby linking site types which are related thematically, eg. all engineering and

manufacturing sites are grouped under INDUSTRIAL. Although the British standard includes the concept of class, the broadest noun term is the Top Term. This thesaurus does not have Top Terms as the classes are not part of the hierarchy.

2.1 Relationships

There are three basic relationships within a thesaurus. These are:

the *Equivalence* relationship

the *Hierarchical* relationship

the *Associative* relationship

To create the thesaurus these relationships were applied to each term.

2.1.1 The Equivalence relationship

This is the first relationship to be decided. A term can be "preferred" or "non-preferred", meaning that a preferred term is the term that will be used in the hierarchies and will be the term used for indexing. A non-preferred term is a term that has the equivalent meaning to the preferred term but is not used for indexing. This might be because the term is:

a) a Synonym

eg. Engineering Works USE **ENGINEERING FACTORY**.

ENGINEERING FACTORY is the more accepted term and so is the preferred term whilst Engineering Works is a variation and so is used as a pointer towards the preferred term.

b) a Quasi-Synonym

eg. Hall Of Memory USE **COMMEMORATIVE MONUMENT**

Where a term is treated as a synonym within a particular subject area.

2.1.2 The Hierarchical relationship

The second stage is to group the preferred terms into hierarchies. They are first gathered into conceptual groups, for example all types of Commemorative Monuments. Then within each conceptual group the terms are further divided into levels going from the broadest type of term to the narrowest and most specific type of term.

eg.	**COMMEMORATIVE MONUMENT**	Conceptual group
	COMMEMORATIVE STONE	Broadest level or BROADER TERM
	DATE STONE	Narrowest level or NARROWER TERM

Here the terms **COMMEMORATIVE STONE** and **DATE STONE** are both types of Commemorative Monuments but **DATE STONE** is a more specific form of **COMMEMORATIVE STONE** so can become a narrower term of it.

A thesaurus can be poly-hierarchical. That is to say, a broad term can appear in more than one hierarchy and under more than one class.

eg. **FACTORY**
EXPLOSIVES FACTORY

EXPLOSIVES MANUFACTURING SITE
EXPLOSIVES FACTORY

EXPLOSIVES FACTORY appears under two separate hierarchies.

In the creation of hierarchies it is sometimes necessary to use a term to group archaeological event types together but that grouping term itself is not intended to be used to index with. This is referred to as a non index term and is identified in the attached listings as a non-bold, capitalised term (eg. AIR DEFENCE SITE) whilst an index term is identified as a bold, capitalised term (eg. **ANTI AIRCRAFT BATTERY**).

2.1.3 The Associative relationship

Terms can be associated with each other but not necessarily connected by a hierarchy. This means that one site type can be associated with another which comes under a different broad term but where the two site types are similar in concept. These are referred to as "related terms". Such terms are often used as an aid to help enquirers find terms similar to the initial term which are not always immediately obvious.

eg. **CORDITE FACTORY**

RT **GUNCOTTON FACTORY**

A **CORDITE FACTORY** is similar to a **GUNCOTTON FACTORY** and vice versa, so the related term is another term that should be looked at if the enquirer wants to broaden their original search.

2.2 Class

The terms within the thesaurus are grouped by classes and not the broadest noun term (Top Term). These groupings have been used to aid search and retrieval but are not part of the hierarchy of terms.
Site types are included in a class on the basis of the criteria set out in the class definitions. Within each class, groups of broader terms can be used to further sub-divide terms. These broader terms reflect the overall conceptual framework of the thesaurus.

2.3 Scope Notes

Scope notes are the final part to be added to a term. A scope note provides a clear indication as to exactly how the term is to be used in the context of this thesaurus. That is, it will provide a definition and any point that should be borne in mind for the use of the term, eg.

COMMUNICATION TRENCH

SN A trench, usually linking two or more rows of trenches, enabling the conveyance of messages or equipment safely from one trench to another.

From the definition it is obvious that this is trench used to enable communication between rows of trenches.

3. Rules for vocabulary control

The rules that have been adopted regarding the choice and form of terms within this thesaurus are as follows:

Synonyms

The thesaurus controls the use of synonyms and quasi-synonyms to improve indexing and retrieval, by the use of preferred and non-preferred terms. Where non-preferred terms have several meanings, there can be more than one preferred term and guidance on their use may be given by a scope note.

Homographs

The use of homographs (words with the same spelling but different meanings) has been restricted within the thesaurus. eg.

PILLBOX

PILLBOX (SHELLPROOF)

PILLBOX (VARIANT)

Singular or Plural

Site types appear in the singular; a site type will only appear in the plural if the plural is the common usage.

Punctuation

Punctuation has been omitted from the hierarchical and alphabetical lists within the thesaurus as its inclusion inhibits retrieval. However, it has been retained within the scope notes to ensure that the definition is understandable.

Spelling

Spelling follows *The Shorter Oxford English Dictionary* (Third Edition 1986), apart from rare exceptions where common practice in the field of archaeological and architectural recording differs from this.

Hyphens

Hyphens are not used in the thesaurus as their inclusion inhibits retrieval. Therefore hyphenated words are treated as two words.

Compound Terms

Complex compound terms are divided up into single concepts, except where this affects the meaning, or where the use of such a term is well established, eg. **SEAPLANE.**

Multiple Indexing

It is common practice when indexing, to assign as many thesaurus terms to each item as are necessary, to express all aspects of the concept. Using this thesaurus it would be possible to index a record for a multi phase site or structure with terms that relate to each phase of the monument.

Language Order

Natural language order is used for all preferred and non-preferred terms eg. **MILITARY BASE**, not **BASE, MILITARY**.

Alphabetisation

Word-by-word alphabetisation is used throughout the thesaurus.

Abbreviations and Acronyms

Abbreviations and acronyms have been omitted from the thesaurus, eg. use **PRISONER OF WAR CAMP** not POW Camp.

Loan-words/Foreign and Classical Terms

Terms which are well established within the English language, or are in common use within the archaeological or architectural community, are included within the thesaurus.

4. Using the Thesaurus for indexing

Good indexing policies and a commitment to improving the quality of indexes are central to the successful operation of the thesaurus on computerised databases. The following guidelines are suggested to obtain maximum advantage from the use of the thesaurus.

a) Validation

The validation of indexing terms as they are entered on to a database is one of the most effective forms of vocabulary control and of increasing retrieval from the database. The thesaurus serves as a master vocabulary file to check the indexing terms used by indexers and searchers. The system can reject non-preferred terms and, if desired, the preferred terms can be automatically substituted, except where there is more than one alternative. A browsing facility can easily lead the indexer to valid terms in a broad, or more restricted, subject area. In addition, a facility for proposing candidate terms can allow users to index records temporarily with a term not at present included in the thesaurus (See 7. Updating and Maintenance below).

b) Recording Practice Guidelines

It is recommended that sections on indexing policy reflecting the requirements of the system's end-users are included in the Recording Practice Guidelines for the database, together with instructions for the use of the thesaurus.

c) Levels of Indexing

The thesaurus is designed for use at the most specific level of information available at the time of indexing. Indexers should therefore use the most specific term (ie narrow term) appropriate for indexing. The detail to which multidisciplinary events should be indexed will reflect user requirements and available resources. The thesaurus allows a flexible approach as it places no restrictions on what may constitute an event for any particular site.

5. Using the Thesaurus for Retrieval

The thesaurus is specifically designed to assist users in maximising the retrieval of information from a database. The hierarchical nature allows the user to retrieve information at different levels or by different concepts according to their needs. By structuring queries in different ways, eg. to include (or exclude) records indexed with narrow terms or records indexed with related terms or with both narrow and related terms, it is possible to expand or contract the information retrieved.

Full guidance on retrieval and the use of the thesaurus should be included in any user guide for a system. It may also be helpful for users to have an alphabetical listing of terms with the number of occurrences on the database. This information will assist users in making enquiries at the appropriate level for their needs, and should be updated regularly.

The thesaurus is closely linked to indexing and retrieval needs and its effective application will benefit from the monitoring of enquiries to the database and the efficiency of retrieval. The recording of enquiries and retrieval problems, together with their regular review, should therefore help to improve the Thesaurus and the indexing of the database.

This thesaurus covers terms for site types but will frequently be most effective when used with other database fields with controlled entries, eg. Period or Date, to refine the search. Clear guidance on such fields, their use in combination with the thesaurus and examples of effective searching techniques, should be included in any user guide.

6. The use and future development of the Thesaurus

The thesaurus has been developed using ORACLE database software and is one of the thesauri within the English Heritage AMIE (Archives Monuments Information England) database used by English Heritage and some Historic Environment Records (HERs). The level of detail included in the thesaurus reflects that which is considered by English Heritage to be appropriate for recording archaeological and architectural site types at a national level, based on the current indexing requirements of the

databases held by them. It is recognised that greater levels of detail may be desirable at a local level and where users have a more specialist interest in a particular area of vocabulary. Such requirements will be reviewed as necessary and appropriate action taken, particularly where data exchange may be involved at a national level. The thesaurus can provide rules and a broad term structure which could form a basis for a more detailed linked vocabulary for use in specialised projects or to meet local requirements.

Source: English Heritage, Data Standards Unit, February 2014

Glossary

BROADER TERM (BT)

A term that represents a parent to a term or other terms within a CLASS. The Broader Term (BT) is super-ordinate to its subordinate NARROWER TERM (NT). The relationship between a broader term and a narrower term is usually generic. One term may have many narrower terms, and in turn, each narrower term may itself have narrower terms, thus allowing the thesaurus to be MULTI-LEVEL:

eg. **COMMEMORATIVE MONUMENT** is the broader term of **COMMEMORATIVE STONE**, which is the broader term of **DATE STONE**.

CANDIDATE TERM

A Candidate Term is a new term which has been proposed by users for inclusion in the thesaurus. Each term will be reviewed by the Data Standards Unit and a decision will be made as to whether the term should be included as a PREFERRED or NON-PREFERRED TERM and placed into the thesaurus accordingly.

CLASS (CL)

The highest term within a HIERARCHY. These terms are used merely as grouping terms to aid retrieval and as such are NON-INDEX TERMS.

GENERIC RELATIONSHIP

The principal link between a CLASS or a BROADER TERM and its members, or NARROWER TERMS. This relationship follows the 'all-and-some' rule as seen below:

TRAINING SITE

SOME ↓ ↑ ALL

FIRING RANGE

The diagram shows that *some* TRAINING SITES are **FIRING RANGES**, but *all* **FIRING RANGES** are by their very nature TRAINING SITES.

HIERARCHY

An arrangement of terms showing the Broader-Narrower relationships between them.

HOMOGRAPHS

Homographs (or Homonyms) are terms which have the same spelling but different meanings. In this thesaurus these are distinguished by a qualifier in rounded brackets, eg. **AIRCRAFT HANGAR (TRANSPORTABLE)** and **PILLBOX (VARIANT)**.

INDEX TERM

A term that can be used to describe a site type in records on a database, eg. **DRILL HALL**. In this thesaurus, INDEX TERMS appear in upper case, bold type.

MULTI-LEVEL

A thesaurus structure with varying levels of BROADER and NARROWER TERMS.

NARROWER TERM (NT)

A term that represents a child to other terms within a CLASS; eg. **CANNON BORING MILL** is a Narrower Term of **BORING MILL**. A Narrower Term can have more than one BROADER TERM (BT), eg. **CANNON BORING MILL** is also a Narrower Term of **ARMAMENT MANUFACTURING SITE**.

NON-INDEX TERM

A Non-Index Term (or Guide Term) is a PREFERRED TERM, which cannot be used as an INDEX TERM, but is useful in the thesaurus as a grouping term for retrieval purposes only, eg. RAILWAY ENGINEERING SITE. Non-Index Terms are distinguished in this thesaurus by appearing in upper case, non-bold type.

NON-PREFERRED TERM

A Non-Preferred Term is a term which cannot be selected for indexing or retrieval (eg. it is synonymous with a term which is already in the thesaurus), but which is retained in the thesaurus to point the user to a PREFERRED TERM which should be used, eg. Bronze Plaque USE **PLAQUE.**

POLYHIERARCHY

A POLYHIERARCHY allows a PREFERRED TERM to belong to more than one CLASS or to have more than one BROADER TERM.

PREFERRED TERM

A term which can be selected for retrieval within the thesaurus. Preferred Terms can be INDEX or NON-INDEX TERMS. Preferred Terms appear in upper case within the thesaurus.

RELATED TERM (RT)

A RELATED TERM is a PREFERRED TERM which can be linked to another PREFERRED TERM conceptually but not hierarchically, eg. **ORDNANCE FACTORY** and **ARSENAL**. The thesaurus allows for terms to be related in the *same* hierarchy when a particularly strong link occurs.

SCOPE NOTE (SN)

A limited definition of a term and/or guidance on its use.

SYNONYM

A term having a different form/spelling but the same or nearly the same meaning as another term, eg. Sound Dish and **SOUND MIRROR**.

UPWARD POSTING

The treatment of NARROWER TERMS as if they are equivalent to, rather than a species of their BROADER TERMS. Upward posting is used where the level of detail, suggested by a term is considered too specific for the thesaurus, eg. Rifle Factory USE **ORDNANCE FACTORY**.

USE

USE indicates the PREFERRED TERM which should be used for a NON-PREFERRED TERM, eg. Gunpowder Mill USE **GUNPOWDER WORKS**.

USE FOR (UF)

USE FOR usually abbreviated to UF, indicates the NON-PREFERRED TERM(S) covered by a PREFERRED TERM;

eg. **ENGINEERING WORKS**

UF Engine Manufactory

Engine Works

Traction Engine Works

WORD-BY-WORD ALPHABETISATION

The alphabetisation of the terms within the alphabetical list of the thesaurus follows the word-by-word format whereby terms are listed alphabetically by word as opposed to letter-by-letter. See example below. In the word-by-word format, a space is alphabetized before any letters or numbers. For example, "BUS STOP" would come before "BUSH." In a letter by letter sort, the spaces between words are ignored, so "BUSH" would come before "BUS STOP."

BUS STOP, BUS STATION, BUST, BUS TERMINAL, BUSH

Word-by-word	**Letter-by-letter**
BUS STATION	BUSH
BUS STOP	BUS STATION
BUS TERMINAL	BUS STOP
BUSH	BUST
BUST	BUS TERMINAL

Source: English Heritage, Data Standards Unit