



ENGLISH HERITAGE



Management of Research Projects
in the Historic Environment
MoRPHE Project Planning Note 2
Developing Controlled Vocabularies

Version	Date	Comments
1	May 2006	First published version

Preface

MoRPHE Project Planning Notes form an integral part of the MoRPHE project management methodology. The Project Planning Notes provide specific guidance on the *management* of particular types of project, not guidelines on the techniques used.

They are intended to be presented together with, and read with '*The MoRPHE Project Managers Guide*' which gives generic guidance on project management.

1.0 Introduction

For the purposes of this guide the term 'controlled vocabulary' covers a variety of forms. These include classification schemes, simple wordlists and more complex polyhierarchical thesauri. The various types of controlled vocabulary are further explained in the Glossary.

This guide is intended to be used by managers and members of Historic Environment project teams who are following the MoRPHE project guidance and want to develop a controlled vocabulary.

Controlled vocabularies are used to index information held in information systems such as historic environment records, context record databases or museum collection catalogues. Use of a controlled vocabulary improves the accuracy of information retrieval. Their use may also be a requirement of national and/or internationally recognised data standards.

The development of a controlled vocabulary usually forms part of a larger project but may be developed in isolation. In either case the development can be said to be a project in itself and this planning note assumes that the development will be run as a project in its own right, even if it is a project within a project.

This project planning note covers the **management** processes and products involved in the development of controlled vocabularies.

2.0 Planning

This section provides advice on the likely stages involved in the development and implementation of a controlled vocabulary.

2.1 Setting objectives

The requirement for a controlled vocabulary will normally come from an existing project or programme of work. Commonly this is the need for a terminology to allow the recording of information in a database for example, to provide a controlled list of names. However it could also be driven by a perceived future need and as such it may initially be created in isolation before being incorporated into the system.

It is essential that the scope of the project is outlined at the start and that the field of knowledge which it is intended to cover is well defined. This will ensure that the vocabulary is fit for purpose and doesn't become increasingly complex.

Usually the inclusion of terms in the controlled vocabulary will be restricted to a particular concept such as people, places or objects and it may only contain terms found within a specialised subject area or field of knowledge.

The important points to remember are:

- What is the scope of the controlled vocabulary?
- Who is likely to use it?
- Will it be available to a variety of audiences or only for internal use?
- Format for dissemination. How will it be used?

2.2 Useful techniques for estimating time and budget

Once the scope of the vocabulary has been identified the next stage is to identify whether there are any similar products available which could be used as a framework on which to build the vocabulary.

The development of a controlled vocabulary can be an expensive and resource intensive process and wherever possible the use of pre-existing vocabularies should be considered.

If no existing vocabulary meets the requirements then the project will need to factor in the time and budget for the research which will need to be done to identify the terms to include in the vocabulary. Often this research will simply involve talking to the users and compiling the list of the terms they use in their work.

If the controlled vocabulary will take the form of a thesaurus and include over 500 terms then more research time will be needed. As a rule of thumb, you should initially allow at least 20 minutes per term.

The time spent on researching terms can be reduced, depending upon the complexity of the vocabulary, by assigning the work to a member of the project team who has a good working knowledge of the field to be covered. Alternatively the work could be contracted out to a domain specialist, although this will obviously impact on any budget.

2.3 Risks and their management

When developing controlled vocabularies it is very difficult to factor in risk. This is mainly due to the dynamic nature of the terminology in use. However the following risks should always be considered:

Scope creep

The controlled vocabulary increases in scope to include subject areas not initially identified. This will impact on both the budget and timescale as more research will be needed to ensure the correct terminology is included.

Increasing the number of terms in the vocabulary

With large and complex vocabularies it is nearly impossible to say at the outset what the final number of terms to be included will be. As such, a flexible approach to defining timescales and budgets should be taken. Any budget should include a small amount set aside as a contingency fund for unexpected work which may arise.

Never underestimate how long will be required to develop your vocabulary.

CASE STUDY

The UK Archival Thesaurus (UKAT) was developed in 2003 to provide a consistent approach to subject indexing within the UK archive community. It was based on a pre-existing thesaurus (UNESCO thesaurus) and during the start-up stage of the project it was anticipated that there would be no more than 3000 candidate terms to add to the existing 6000 already in the UNESCO thesaurus. However, the final number of candidate terms received over the course of the 18 month project, exceeded 13000 all of which had to be included during the project life cycle.

This was mainly down to two factors. Firstly inadequate research was undertaken prior to the start of the project to identify likely sources for terms and secondly an increased awareness within the community led to more archives coming forward and providing their own terminologies for inclusion.

Lack of knowledge in the subject area

As with scope creep, an inadequate understanding of the subject area can lead to underestimating the timescale and budget required to complete the project. Work undertaken during the planning stages and the initial research stages can counteract this by ensuring that all possible sources of terminology have been identified.

Inappropriate type of vocabulary

Selecting the correct type of vocabulary from the outset will impact on the timescale and budget. There is no point in budgeting for a flat wordlist of 100 terms when the user requirement is a fully implemented thesaurus of 1000 terms. The opposite also applies: Why spend time and money on a thesaurus when a simple look-up list will meet user needs

Lack of editing tools

Although it is possible to develop complex thesauri using pen and paper it is simpler and more efficient to use a software application, where one exists. It may be that such a resource already exists within the organization but if not then serious consideration should be given to the procurement of relevant software as part of the project.

CASE STUDY

The HEREIN thesaurus is a multi-lingual thesaurus developed for the European Heritage Network. Initially conceived in 1999 it took six years to reach the point where it was made available on the internet. This delay was due to problems with the procurement of suitable thesaurus management software and as such most of the work was carried out using paper and word processing software. The delay could have been minimized and the workload reduced if software had been developed prior to the development of the thesaurus.

2.4 Likely Products and outcomes

This section shows a number of the main products that may result from a project to develop a controlled vocabulary. Figure 1 provides a breakdown of likely products from a project of this sort.

- Thesaurus maintenance software – either developed in house or bought off-the-shelf
- An introduction or guidance manual explaining why the vocabulary has been developed and what it covers.
- Finished vocabulary – Depending on which type of vocabulary has been developed this may be a simple list circulated either digitally or in paper format or if it is to be a printed thesaurus it may require a publication schedule and budget.
- Peer review questionnaire
- Integrated vocabulary – The vocabulary may require integration into a database. This usually takes the form of a look-up list associated with a particular database field or a series of forms which can be used to interrogate the vocabulary.
- Candidate term submission procedure
- Promotion and awareness-raising

2.5 Likely Project Stages

The project proceeds through a sequence of stages.

Start-up and Initiation

A project proposal should outline why the controlled vocabulary is needed, what the scope of the project will be and the expected outcomes and benefits to the organisation.

This stage allows you to assess whether the project should be implemented. If a substantial project is proposed, then early review (Review Point R1) may suggest the need for additional consultation and planning may in a separate Initiation stage, leading to further Review (Review Point R2) and formal approval for the project.

Execution Stages

Once it has been agreed that it is worthwhile to undertake the project then project execution can begin. The following execution stages are appropriate

- Research stage – in which the terms to be included in the controlled vocabulary are identified and defined.
- Software Procurement
- Development stage – when the terms are brought together and developed into a structured wordlist. In a wordlist this may simply involve placing the terms in alphabetical order. In a more complex classification scheme or thesaurus this will involve grouping terms into common areas and identifying relationships between them.
- Peer Review – When the controlled vocabulary has been developed it should then be circulated to the users and other interested parties for them to assess whether it meets the requirements.
- Dissemination and archiving – Having assessed the vocabulary, and once it has been signed off, then it can be made available for use. A copy of the vocabulary should be archived.
- Closure – Once the vocabulary has been disseminated then the project can be closed.

The continued development of the vocabulary falls outside the scope of the initial development project. However proposals for its ongoing maintenance should be planned and agreed as a product of the development project.

- Post-project evaluation – An evaluation phase should start after the vocabulary has been in use for an agreed period of time. This will highlight any problems in using the vocabulary and any areas which have been insufficiently covered in the initial development. This phase usually falls outside of the lifecycle of the project but should be undertaken to determine whether the project was successful.
- Ongoing maintenance – Controlled vocabularies evolve with use; they are constantly being updated, enlarged and adapted to reflect changes in the field of knowledge that they cover. As such it is essential that the vocabulary is maintained and kept up-to-date. An owner should be identified to whom the ongoing maintenance can be assigned. Preferably this should be someone familiar with both the development of controlled vocabularies and the field of knowledge that it covers.

In addition it is useful to establish a small working group which can advise on future developments and review any candidate terms which have been submitted by the user community.

3.0 Project execution

The execution stages focus on the actual work involved in delivering the controlled vocabulary. The list of products used to develop the controlled vocabulary is illustrated in Figure 1. The order in which they might be developed is shown in Figure 2.

3.1 Research Stage

Having defined the scope, and type, of the controlled vocabulary the first stage should focus on identifying resources which can be 'mined' for terms to be included.

In addition to bibliographic sources there are numerous pre-existing wordlists, thesauri and classifications schemes which include terms covering many of the aspects of the cultural heritage, including:

- Getty Art and Architecture thesaurus
- British Museum Materials thesaurus
- NMR Thesauri
- Office of National Statistics lists of English counties, districts and parishes
- UNESCO thesaurus
- UKAT thesaurus
- HEREIN thesaurus
- Library of Congress Subject Headings
- Dewey Decimal Classification
- Universal Decimal Classification

Not only do these represent a rich source of terms, they may also provide you with a basic framework on which to build your vocabulary. The adaptation of existing vocabularies is a cheap and effective way to begin the complex task of defining relationships between terms. This is particularly true for thesauri which rely upon the editor having a good understanding of the fundamental principles of thesaurus construction.

Unless the vocabulary being developed is a simple wordlist, this research stage will continue throughout the life-span of the product. Although, it is expected that at the end of the project a fully-developed vocabulary will exist which fulfils user-needs and can be made available, experience shows that vocabularies are dynamic and constantly evolve through use.

This usually means that mechanisms need to be put in place which allow for their continued development after the project life-cycle has been completed.

3.2 Development stage

The length of this stage is dependent upon the type of vocabulary being developed. A simple wordlist comprising only a handful of terms can be researched, developed and implemented in a matter of hours, particularly if the editor is familiar with the subject area and the user requirement is clearly defined.

However a thesaurus or classification scheme will require considerably more time to develop as the complex relationships between terms need to be defined to ensure that the logical structure is consistent and conforms to the recognized standards

3.3 Skills required

Whilst simple wordlists can be, and indeed are, compiled quickly by anyone, more complex controlled vocabularies require a greater range of skills.

Attention to detail and an understanding of the subject area as well as a methodical approach to problem solving are all essential qualities required of the developer. A high level of literacy and a good understanding of words and word-origins are pre-requisites. Good research skills are also useful.

English Heritage has a dedicated unit, the Data Standards Unit, who are experienced in all aspects of controlled vocabulary creation. They provide advice and assistance to anyone working in this area.

The Forum on Information Standards in Heritage (FISH) is a consortium of heritage bodies and organizations working together to promote the use and development of data standards in the wider Heritage community.

FISH also maintains INSCRIPTION, a collection of controlled vocabularies providing various terminologies for indexing all aspects of the built and buried heritage.

MDA is the UK's lead voice for documentation and the management of information about museum collections. They have experience in setting professional standards and helping museum professionals maintain them.

For contact details and further information see section **6.0 Further Information**

4.0 Review

The review of a finished vocabulary is, to a certain degree, subjective. Whilst it is possible to ensure that the quality of the actual list is measurable in terms of, for example, the number of terms included or whether the terms are spelt correctly, it is far more difficult to be objective about the quality of the nature of the relationships between terms or the hierarchies. The developer can only ensure that the vocabulary complies with the necessary standards and guidelines and is 'fit-for-purpose'.

4.1 Relevant standards and guidelines

At present there are a range of standards available for the construction of controlled vocabularies. However these are in the process of being revised and will be superseded by a new British standard which will ultimately comprise 5 parts. Parts 1 and 2 have already been developed and are listed below.

Structured vocabularies for information retrieval — Guide — Part 1: Definitions, symbols and abbreviations / British Standards Institution. - London : BSI, 2005. (BS 8723-1:2005) – ISBN 0 580 46798 8.

Structured vocabularies for information retrieval — Guide — Part 2: Thesauri / British Standards Institution. - London : BSI, 2005. (BS 8723-2:2005) - ISBN 0 580 46799 6.

Parts 3, 4 and 5 will be published in the near future and will focus on:

- Vocabularies other than thesauri
- Interoperation between multiple vocabularies (with multilingual thesauri as a special case)
- Interoperation between vocabularies and other components of information storage and retrieval systems.

In addition to the standards the following guide is also recommended:

Thesaurus construction and use: a practical manual / Jean Aitchison, Alan Gilchrist, David Bawden. - 4th ed. - London : Aslib, 2000. ISBN 0-85142-446-5

4.2 Approaches to assessment of quality

This will usually take the form of a peer review, where a group of reviewers with the required knowledge are asked to comment on the vocabulary. It is circulated along with a questionnaire and the reviewer is asked to comment on whether:

- the project has achieved its objectives
- the scope of the vocabulary is relevant and 'fit-for-purpose'
- the vocabulary conforms to the recognized standards

At the end of the peer review period, the questionnaires will be returned to the project team and any comments noted.

If the peer review highlights errors and/or omissions it may be necessary to make further amendments to the vocabulary before dissemination to a wider audience.

5.0 Archive & Dissemination

The archive for this sort of project should aim to ensure that decisions taken during the development of the controlled vocabulary (for example why particular terms or groups of terms were included or excluded) can be reviewed at a future date, for example if a further edition is planned. Generally this information should be recorded in the project Issue Log. The returns from peer review should also be retained to assist future editing.

Dissemination of the vocabulary is dependent upon its intended audience and its intended use.

If the vocabulary is to be integrated into a database then it may only be made available to users through the system and it may not be necessary to publish digital or paper copies.

However if the vocabulary is to be used by external partners or organizations then it may be necessary to provide it in a variety of formats including digital and paper.

Also it should be noted that if the vocabulary is a thesaurus and is intended as a traditional publication then a copy should be deposited with Aslib, the Association for Information Management, in the UK and, if appropriate, with the international clearinghouse in Toronto. The address of which is as follows.

Subject Analysis Systems Collection
Faculty of Library Information Science
University of Toronto
140 St. George Street
Toronto
Ontario M5S 1A1
Canada

6.0 Further Information

Useful Contacts

Aslib, The Association for Information Management

Holywell Centre
1 Phipp Street
London
EC2A 4PS
Tel: +44 (0) 20 7613 3031
Fax: +44 (0) 20 7613 5080
Email: aslib@aslib.com

Data Standards Unit

National Monuments Record
Kemble Drive
Swindon
SN2 2GZ
Tel: +44(0)1793 414700
Fax: +44(0)1793 414770
Email: dsu.info@english-heritage.org.uk

MDA

The Spectrum Building
The Michael Young Centre
Purbeck Road
Cambridge
CB2 2PD
Tel: +44(0)1223 415760
Fax: +44 (0)1223 415960
E-mail: mda@mda.org.uk

Useful websites

Forum on Information Standards in Heritage

<http://www.fish-forum.info/>

INSCRIPTION

<http://www.fish-forum.info/inscript.htm>

National Monuments Thesauri

<http://thesaurus.english-heritage.org.uk/>

7.0 Acknowledgement

Particular thanks are due to the following who participated in a review of this project planning note: Mark Dunkley; Jaqui Huntley; Martin Newman; Jane Siddell; Sue Stallibrass; Richard Whittaker.

8.0 Contact details

This Project Planning Note was authored by Philip Carlisle, English Heritage Data Standards Unit. Email Philip.carlisle@english-heritage.org.uk. Comments to assist future versions are welcome.

Glossary

Classification Scheme

A method of organizing terms according to a set of pre-established principles, usually characterized by a notation system and a hierarchical structure of relationships among the entities. Unlike a thesaurus a classification schemes does not establish the equivalence or associative relationship between terms.

Controlled Vocabulary

A limited set of terms or notations in a thesaurus or classification scheme that must be used both for indexing and searching. In a controlled vocabulary consisting of terms taken from a natural language, the use of synonyms and homonyms is avoided for terms recommended for use in indexing. In most cases some structure is imposed on the terms and notations so that those whose meanings are related are linked in some way.

Polyhierachical Thesaurus

A thesaurus in which a term can be assigned to more than one hierarchy. For example, in a thesaurus dealing with animals, the term 'DOGS' could be assigned to both a 'WORKING ANIMALS' hierarchy and a 'PETS' hierarchy. Although more complex than a normal thesaurus, a polyhierarchical thesaurus reduces the need for the user to be able to identify the correct hierarchy.

Term

A term used consistently in indexing to represent a given concept. In a thesaurus these can be either Preferred (ie. used for indexing) or Non-preferred (eg. synonyms used as lead-in terms).

Thesaurus

A structured list of terms which have semantic relationships defined between them. The terms are grouped into hierarchies and the equivalence relationship (eg. synonyms, quasi-synonyms) and associative relationship (conceptually similar terms which are not linked hierarchically), are established.

Wordlist

A flat, alphabetical list of terms used for indexing. Unlike a thesaurus a wordlist does not define relationships between terms.

Fig 1. Product breakdown structure of the specialist products created for a controlled vocabulary

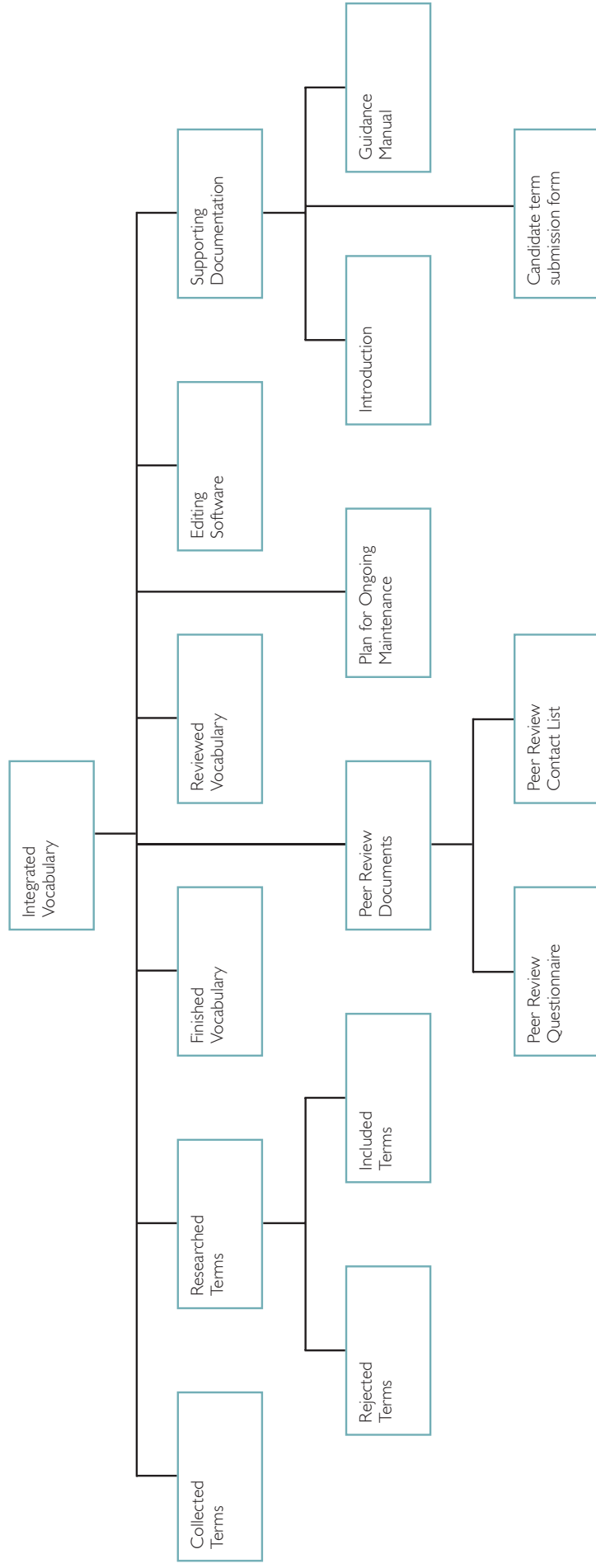


Fig. 2 Product flow diagram of the specialist products created for a controlled vocabulary

