Practical approaches to standardizing vocabularies: the Cultural Heritage experience

Philip Carlisle[a *]
[a] English Heritage, National Monuments Record
Kemble Drive, Swindon SN2 2GZ
Wiltshire, United Kingdom
[*] philip.carlisle@english-heritage.org.uk

Abstract: The European Heritage Network was established as part of an EU funded project (HEREIN) which ran from 1998-2003 with the aim of creating a pan-governmental network for the Cultural Heritage Sector. One of the key deliverables of the project was a multilingual thesaurus of Cultural Heritage Policy Terminology (the HEREIN thesaurus) to allow the indexing of the National Heritage Policy reports of the EU member states in the two official languages of the Council of Europe (English and French) and Spanish. This paper will outline the practical approaches taken to develop the thesaurus and its extension from the original three languages using the HEREIN thesaurus as a case study.

Keywords: Cultural Heritage; thesauri; indexing; international standards; multilingual; European cooperation; mapping.

1. Introduction

The European Heritage Network brings together government departments and agencies responsible for cultural heritage under the umbrella of the Council of Europe. It was established in 1999 as part of the HEREIN project following the 4th European Conference of Ministers responsible for the Cultural Heritage held in Helsinki in 1996 which recommended that the Council of Europe "consider setting up a permanent information system for the benefit of national authorities, professionals, researches and training specialists in touch with heritage developments in other countries" [01]. The network has since become a reference point for government bodies, professionals, research workers and non-governmental organisations active in this field and currently includes representatives from 41 countries.

The website of the network [02] provides a database of the national heritage policies of the EU member states as well as a multilingual thesaurus allowing users to access the policies of other countries using their own language.

This paper will outline the development of the thesaurus and the methodology and approaches used to provide a practical, multilingual tool using the Thesaurus of Cultural Heritage Policies (HEREIN thesaurus) as a case study.

2. Standardizing terminology

With the increasing availability and use of the internet the need to standardize terminologies to facilitate searching has become a major issue for information providers. This is particularly important where the same information is available in multiple languages and needs to be conveyed to users from across a wide geo-political area.

Without standardized vocabularies

2.1. Wordlists

Most people are aware of the existence of controlled vocabularies. The simplest form is the wordlist; an alphabetical list of keywords to be used to control the entry to a database field.

Simple wordlists are limited by their inability to express any relationships between the words contained in the list. In the following list there are four types of building:

Bungalow
Castle
Fort
House

This simple wordlist, although useful, could be given greater meaning by grouping the terms together into hierarchies related by function:

Dwellings
    Bungalow
    House
Defensive Buildings
    Fort
    Castle

This more complex wordlist is already more useful and allows the user to search at the broader level using the "parent" (eg. Dwellings) to retrieve all records indexed with any of the "child" terms or simply to search on the individual child terms.

2.2. Thesauri

The structure of a thesaurus provides even greater flexibility than a complex wordlist. It allows three basic relationships to be established between terms:

- hierarchical - groups terms together according to certain criteria for example, function or type.
- equivalent - allows the user to define which of two synonyms should be used as the "preferred term" when indexing.
- associative – allows terms to be related to one another where no explicit hierarchical or equivalent relationship may exist

In addition to creating relationships a thesaurus allows user to define the terms using "scope notes". These can simply be a dictionary-style definition of the term but they can also include guidance on how the term should be used.

The construction of a thesaurus should conform to the standard as defined in the relevant ISO document. There are currently two standards, one governing monolingual thesauri (ISO2788) [03] and one governing multilingual thesauri (ISO 5964) [04].

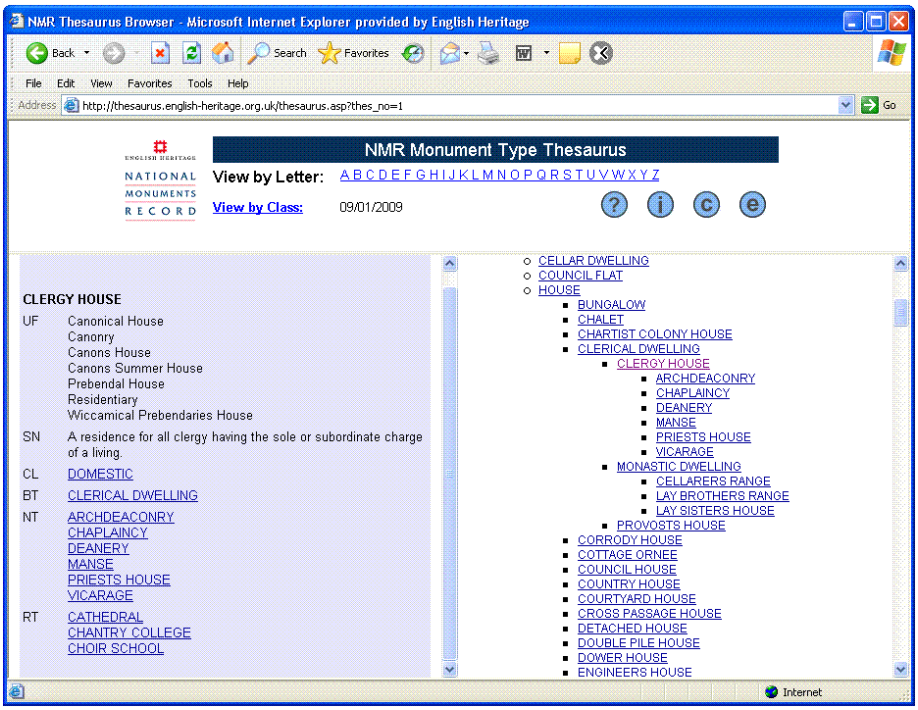

Fig.1 Screenshot of the term CLERGY HOUSE, showing hierarchical and alphabetical displays. Copyright, English Heritage, 1999.

These standards have been in use since the late 1980s and are currently being revised but the basic principles are still valid.

In a display, whether printed or digital, the following conventions are used to show the relationships:

- USE – indicates the "preferred term" to be used where synonyms exists
- UF- indicates the "non preferred term" or "guide term"
- SN – indicates the scopenote or definition
- BT – indicates the Broader Term or "parent"
- NT- indicates the Narrower Term or "child"
- RT – indicates the Related Term

3. Developing the HEREIN Thesaurus

It was recognized early on, that a network such as that proposed by the Conference of Ministers, would need a database which would allow searching across multiple languages, as the main function of the network was to provide access to the national heritage policies of the EU member states. As such, some form of multilingual vocabulary would be required and a working group was established to develop the HEREIN thesaurus.

As the project was being run under the auspices of the Council of Europe it was necessary for each policy document to be produced not only in the native language of the country, but also in either English or French[1] and so those two languages would form the basis of the thesaurus with Spanish[2] being added as the third language.

3.1 Methodology

Before the thesaurus could be developed it was necessary for the thesaurus working group to identify the terms which would be required for the thesaurus. In addition to the documents from France, Spain and the UK the national policies of Norway, Eire and Hungary were chosen to be 'mined' for terms. These terms were then augmented with others derived from various legislation and specialized documentation such as legal dictionaries. This initial phase resulted in approximately 1200 terms in each language.

Many of the terms identified were deemed to be too specific to the particular legislation of one country and so it was decided to limit the initial number of preferred terms to 500 with the more specific terms being denigrated to non-preferred status. Not only would this make the thesaurus easier to manage but it would also ease usability by focussing on more generic terminology common to the 3 languages.

The terms were then grouped into 9 broad categories and a hierarchy was constructed in each of the 3 languages.

1 *Agents (organisations and people)*
General terms for people and organisations involved in heritage, for example local authority.

2 *Heritage Category*
Specific categories and objects connected with heritage, for example protected sites.

3 *Documentation*
The tools and references or standards used in the creation of documentation, for example computerized database.

---

[1] The two official languages of the Council of Europe

[2] Spanish was chosen as the third language as the website was also being translated into Spanish and it was felt that 3 languages would provide a stronger foundation for future expansion of the thesaurus

4 *Legal systems*
Terms used to represent specific legal concepts, for example listed building. It contains policies. The act of carrying out a policy would be in Interventions

5 *Interventions*
Actions carried out related to heritage, for example archaeological excavations It would include the act of carrying out a policy or strategy but not the policy itself which would be in Legal systems

6 *Professional training, skills and qualifications*
Terms for specific skills or professions and training connected to heritage, for example art restorer.

7 *Access and Interpretation*
Terms covering the provision of access to heritage and the interpretation of heritage, for example school trips.

8 *Economic and Financial systems,*
Terms covering finance and finance related activities, for example public grants.

9 *Broad concepts*
General concepts related to heritage, for example archaeology.

As no single language was being used as a source language it was necessary to ensure that each of the hierarchies was exactly equivalent to its two linguistic counterparts. As the structures of the French, Spanish and UK legal and governmental organization differed it was often necessary to create new terms in the other two languages to ensure that the hierarchies matched. This was particularly true for the first group where the loan terms of ‘autonomous communities’ and ‘communautés autonomes’ were created in English and French to allow the inclusion of an essential Spanish term ‘comunidades autónomas’.

Once the hierarchies had been completed each term was then defined in each of the three languages and then translated. These definitions were then compared to define the degrees of equivalence. Five degrees of equivalence as defined in ISO 5964 were used and each language was compared to the other two languages, ie. En to Es and Fr, Fr to En and Es, Es to En and Fr to ensure that the equivalences were true for all three languages:

1 Exact equivalence -where the meaning and scope of the terms are the same in both languages being mapped and both terms are capable of being preferred terms.

2 Inexact equivalence - here a term in the target language expresses the same general concept as the source language but the meanings are not precisely identical

3 Partial equivalence – where the term in the source language has no exact match but a near translation can be made by choosing a term in the target language which has a slightly broader or narrower meaning

4 Single to Multiple – where the term has no exact match but can be expressed by a combination of two or more preferred terms in the target language

5 Non-equivalence – where the target language contains no equivalent concept

Where terms were found to have no equivalents and the term was deemed to be essential, it was agreed that either the source term would be taken as a loan term (ie. in its original form) or translated.

3.2 Mapping and extending the thesaurus

Once the initial tri-lingual thesaurus had been established on paper it was decided to extend it to four languages. Hungarian was chosen as the fourth language as part of the HEREIN 2 extension project.

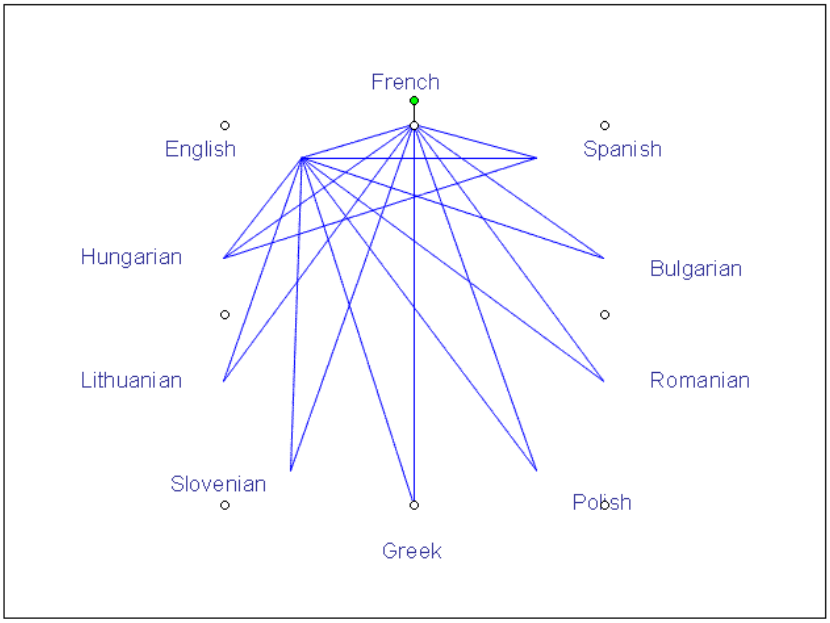

Fig.2. The working thesaurus showing the mappings

Having successfully integrated Hungarian it was then decided to extend to include 6 other languages. These were Bulgarian, Greek (Cypriot), Lithuanian, Polish, Romanian and Slovenian.

To facilitate the extension, software was procured to enable online editing. Unfortunately this was not successful and so new open source software was developed whilst the thesaurus continued to be developed using spreadsheets.

Each new country was also given the option of choosing just 2 of the original languages to map to (see Fig. 2). Inevitably this meant that the new languages chose to map to English and French. This meant that the mapping process would be quicker but didn't accurately reflect the degrees of equivalence between all languages. In fact this form of mapping actually created more problems than it solved particularly where a term in one of the new languages claimed an exact equivalence with term A in English but a single to multiple with term B in French. If term A and B had been defined as begin exactly equivalent then the mapping was not logical.

It was decided in 2005 to go back to basics and to ensure that the software was capable of dealing with any problematic logic loops caused by this and to simplify the mapping process.

The German language was introduced as the new fourth language and the degrees of equivalence were painstakingly reviewed by the thesaurus working group.

Once the software had been loaded with the new core thesaurus, the other languages were reinstalled one by one. The software now checks that the mappings do not create any logic loops and produces a report detailing any errors encountered.

4. Conclusion and future work

Developing a thesaurus from scratch is a resource-intensive, time-consuming process and should not be entered into lightly. Particularly one which involves politics and more than one language! The European Heritage Network is now entering its second decade and work still continues on the thesaurus. It has taken a long time to find a solution which is both practical, financially viable and, with the number of languages now approaching 14, increasingly essential but with the software which is now in place the thesaurus working group are confident that any new countries wishing to join the thesaurus will find the work quicker and easier than those who were there at the start. As well as looking to expand with more languages the thesaurus working group is currently looking at how to handle the problem of synonyms not only within a language but across borders. The Francophone team are trying to solve the problem to ensure that the same concept can be expressed using different terms in French, Belgian, Swiss and Luxembourgeoise whilst still maintaining their national preference.

5. References

[01] IVth European Conference of Ministers responsible for the Cultural Heritage, Helsinki (1996) *Report by the Secretary General submitted in pursuance of paragraph 8 of Resolution (71) 44 of the Committee of Ministers* https://wcd.coe.int/com.instranet.InstraServlet?command=com.instranet.CmdBlobGet&InstranetImage=259999&SecMode=1&DocId=546330&Usage=2

[02] Council of Europe: http://www.european-heritage.net/sdx/herein/index.xsp (2008)
[03] International Organization for Standardization (1986): *ISO 2788: Guidelines for the establishment and development of monolingual thesauri*, 2nd Ed. Geneva: ISO, 1986
[04] International Organization for Standardization (1985a): *ISO 5964: Guidelines for the establishment and development of multilingual thesauri*, Geneva: ISO, 1985