motif

Report on the MoTIF project. Thesaurus construction guidelines and a pilot thesaurus of Irish folklore. **Catherine Ryan**

nli Leabharlann Náisiúnta na hÉireann
National Library of Ireland

dri

Report on the MoTIF project. Thesaurus construction guidelines and a pilot thesaurus of Irish folklore.

Catherine Ryan

Leabharlann
Náisiúnta
na hÉireann
National Library
of Ireland

Digital Repository of Ireland
Taisclann Dhigiteach na hÉireann

DOI: 10.3318/DRI.MoTIF.2014.1

Design: Fidelma Slattery

MoTIF project

MoTIF project team: Sandra Collins, Digital Repository of Ireland; Rebecca Grant, Digital Repository of Ireland; Eoghan Ó Carragáin, National Library of Ireland; Catherine Ryan, Digital Repository of Ireland/National Library of Ireland; Ger Wilson, National Library of Ireland

We gratefully acknowledge the contributions of Dr Jodi Schneider (INSIGHT Centre for Data Analytics, formerly Digital Enterprise Research Institute (DERI)); Dr Elizabeth Mullins (University College Dublin (UCD)); Jane Burns (Royal College of Surgeons in Ireland); Dr Christoph Schmidt-Supprian (Trinity College Dublin); Dr Ríonach Uí Ógáin (National Folklore Collection, UCD); Críostóir Mac Cárthaigh (National Folklore Collection, UCD); and Anna Bale (National Folklore Collection, UCD).

Executive summary

MoTIF (Mo Thesaurus of Irish Folklore) is a collaborative project undertaken by the Digital Repository of Ireland (DRI) and the National Library of Ireland (NLI).

The MoTIF project created a set of guidelines called *Thesaurus construction guidelines: An introduction to thesauri and guidelines on their construction*. Thesauri are indexing and retrieval tools, which improve search and discovery systems and website navigation across all sectors, including cultural heritage as well as business and enterprise. The aforementioned guidelines act as a comprehensive introduction to thesauri and provide guidance on the construction of thesauri using facet analysis, an increasingly popular method of organising terms selected for inclusion in a thesaurus and one recommended by *ISO 25964-1*.

The idea for the project arose following the *Digital archiving in Ireland: National survey of the humanities and social sciences* DRI report, which identified a number of problematic areas in Irish-language names, both personal names and place names, as well a large number of either custom-made vocabularies or international vocabularies adapted for use with Irish content. The guidelines offer advice on how to bridge this vocabulary gap and ensure that professionals have the advice they need to improve their own data procedures by adhering to international standards and best practices such as *ISO 25964-1. Information and documentation: Thesauri and interoperability with other vocabularies*.

In order to demonstrate the principles outlined in the guidelines, a pilot thesaurus of Irish folklore was constructed. This pilot was also called MoTIF.[1] It acts as a sample thesaurus and demonstration of TemaTres[2], the open source thesaurus management software, which users can browse, explore and evaluate for their own vocabularies. The pilot thesaurus is not, however, intended as a complete indexing and retrieval tool.

The Digital Repository of Ireland and the National Library of Ireland will promote international standards and best practices in search and retrieval of content among librarians, archivists and other information professionals in the cultural heritage and enterprise communities by disseminating the guidelines on thesaurus construction and the accompanying pilot to Irish data producers, consumers and custodians.

[1] http://apps.dri.ie/motif

[2] http://www.vocabularyserver.com/

The MoTIF Project

The aim of **MoTIF**, a collaborative project led by the Digital Repository of Ireland (DRI) and the National Library of Ireland (NLI), was to produce guidelines on the construction of thesauri for librarians, archivists, museum professionals and other information professionals. These guidelines act as a comprehensive introduction to thesauri and provide guidance on the construction of thesauri using facet analysis. The guidelines are illustrated by MoTIF, the pilot thesaurus of Irish folklore.

Thesauri are vital and valuable tools in content discovery, and information organisation and retrieval, activities common to all fields, including cultural heritage and higher education as well as business and enterprise. Thesauri allow information professionals to represent content in a consistent manner and enable researchers and the public to find this content easily and quickly. These guidelines will give professionals the advice that they need to improve their own data management processes by adhering to international standards and best practices.

The pilot thesaurus acts as an illustrative guide, providing examples throughout the thesaurus construction guidelines, as well as a sample thesaurus and outlining the international principles and best practices in the guidelines. It also acts as a demonstration of TemaTres, the open source thesaurus management software used in the project, and as a core body of work, which can be expanded further into a complete thesaurus.

Project participants

The **Digital Repository of Ireland** is a national trusted digital repository for Ireland's social and cultural data. The repository will link together and preserve both historical and contemporary data held by Irish institutions, providing a central Internet access point and interactive multimedia tools. As a national e-infrastructure for the future of education and research in the humanities and social sciences, DRI will be available for use by the public, students and scholars.

The Digital Repository of Ireland is built by a research consortium of six academic partners working together to deliver the repository, policies, guidelines and training. These research consortium partners are the Royal Irish Academy (RIA, lead institute); the National University of Ireland, Maynooth (NUIM); Trinity College Dublin (TCD); Dublin Institute of Technology (DIT); the National University of Ireland, Galway (NUIG); and the National College of Art and Design (NCAD). DRI is also supported by a network of academic, cultural, social and industry partners, including the NLI, the National Archives of Ireland (NAI) and Radio Teilifís Éireann (RTÉ). Originally funded from the Higher Education Authority Programme for Research in Third-Level Institutions (PRTLI) Cycle 5 for 2011–15, DRI has also received awards from Enterprise Ireland, Science Foundation Ireland, the European Commission's Seventh Framework Programme (FP7) and The Ireland Funds, and has extended its funding to 2019. The DRI Research Consortium is currently collaborating with a network of cultural, social, academic and industry partners, including the NLI, the NAI and RTÉ.

The mission of the **National Library of Ireland** is to collect, preserve, promote and make accessible the documentary and intellectual record of the life of Ireland and to contribute to the provision of access to the larger universe of recorded knowledge. The NLI offers an exciting programme of exhibitions, events and learning opportunities for people of all ages and interests. It is driving a forward-looking programme of digitisation, digital preservation, and innovative access, visualisation and engagement tools for Ireland's rich cultural collections. For example, the NLI is one of the main contributors to Vufind, an open-source discovery interface, which is used by hundreds of libraries around the world to enhance access to research materials. The NLI also has considerable experience in converting and enhancing metadata for cultural heritage items, including working with linked data resources like Freebase, VIAF and DBpedia.

Introduction to thesauri for information retrieval

Different types of knowledge organisation systems (KOS) and knowledge representation systems abound. From controlled vocabularies to authority lists and from classification schemes to taxonomies, thesauri and ontologies, different systems have been used to define and describe terms and concepts and, in general, to organise knowledge for better search and retrieval for many years.

Specifically, a thesaurus is a controlled vocabulary, which contains hierarchical relationships such as *broader term (*BT) and *narrower term* (NT), equivalence relationships such as *use* (USE) and *use for* (UF) as well as associative relationships such as *related term* (RT). A thesaurus can be referred to as a networked collection of terms where all terms are connected and not only assists users in finding information but also in understanding it.

Improving a user's ability to find the information they are looking for quickly and easily is the main goal of most thesauri and other controlled vocabularies. They are tools that allow both the indexer and the researcher to use the same terms to describe the same subjects or concepts, allowing for easier search and retrieval of information about a particular domain (International Organization for Standardization 2011). By doing so, they support indexing, retrieval, and the organisation and navigation of information.

The relationships in a thesaurus guide users to more general or more specific concepts by allowing them to navigate the vocabulary and choose the most suitable terms for their content. This ability to navigate the thesaurus makes it much more useful than a simple controlled list of terms as it allows a user to browse a subject domain or website. Thesauri also have their place in business. Significant time and money is lost when employees spend time searching for content on an intranet and cannot find it quickly or easily. If a customer cannot find a product on a website, they go elsewhere, usually to a competitor (Stewart 2011). Faceted navigation, the principle of division and arrays are now visible on many websites, especially on those which sell products to consumers.

Project overview

The DRI report *Digital archiving in Ireland* identified a number of problematic areas in Irish language names, both personal names and place names, as well a large number of either custom-made vocabularies or international vocabularies adapted for use with Irish content (O'Carroll & Webb 2012). An opportunity was identified to ensure that professionals have the advice that they need to improve their own data methodologies by adhering to international standards and best practices.

To this end, the MoTIF project produced a set of guidelines, *Thesaurus construction guidelines: An introduction to thesauri and guidelines on their construction* and an accompanying pilot thesaurus, MoTIF, which acts as a sample thesaurus and visual demonstration of the international standards and best practice outlined in the guidelines.

The guidelines are intended as a resource for librarians, archivists and other information professionals who wish to organise and annotate their content for improved search and retrieval according to international standards and best practices including *ISO 25964-1. Information and documentation: Thesauri and interoperability with other vocabularies*. These principles and practices are relevant to the cultural heritage sector as well as to business and enterprise and the advantages of consistent search and retrieval are common across all sectors through the efficient discovery of both cultural heritage content and business products.

The pilot was intended to act as a sample thesaurus and a visual demonstration of the international principles and best practices outlined in that document. It also demonstrates TemaTres, the open source thesaurus management software used to manage and publish the thesaurus, which users can browse, explore and evaluate for their own vocabularies.

Guidelines

A set of guidelines, *Thesaurus construction guidelines: An introduction to thesauri and guidelines on their construction*, were produced to provide a comprehensive introduction to thesauri and advice on how to construct thesauri following international standards and best practice.

A comprehensive literature review was undertaken covering the main aspects of thesaurus construction from both a theoretical and practical viewpoint. The main literature and standards included *ISO 25964: Information and documentation, thesauri and interoperability with other vocabularies* as well as writings by experts in knowledge organisation and management. A full bibliography has been provided in the guidelines document.

The common elements and processes outlined in the literature were collated and brought together in the guidelines to illustrate best practice and international standards and recommendations.

The guidelines cover the basics of thesaurus construction. This included the main elements of a thesaurus:

- Terms and concepts;
- Equivalence, hierarchical and associative relationships;
- Notes; and
- Node labels and arrays.

The *Thesaurus construction guidelines* also outlined the broad steps involved in the thesaurus construction process. Construction is an iterative process in parts and some of the following steps will overlap:

- Selection and recording of terms;
- Determining the structure and display;
- Vocabulary analysis;

- Creating relationships and notes;
- Creating an alphabetical list from the systematic display;
- Review by experts;
- Documentation, including an introduction and editorial guide.

The guidelines also explained facet analysis, an increasingly popular method of organising terms selected for inclusion in a thesaurus and one recommended by *ISO 25964-1*. This was the method used to construct MoTIF, the pilot thesaurus of Irish folklore.

In addition to thesaurus construction and facet analysis, the guidelines also cover the basics of thesaurus planning, what aspects of thesaurus construction and maintenance need to be considered before and after construction. Multilingual thesauri, the process of mapping different thesauri to each other and the relationship between thesauri and the Semantic Web were also briefly touched upon.

Pilot thesaurus of Irish folklore[3]

MoTIF, the pilot thesaurus of Irish folklore, was intended to accompany the *Thesaurus construction guidelines: An introduction to thesauri and guidelines on their construction* and act as a sample thesaurus and a visual demonstration of the international principles and best practices outlined in that document.

There are a number of differences between a pilot and a full thesaurus. As a pilot, MoTIF is not intended as a complete indexing and retrieval tool but instead as a visual tool to assist readers to understand the principles, techniques and elements outlined in the guidelines document itself. The terms and concepts, relationships and notes present in the pilot are presented as examples and should not be considered exhaustive.

The project used the open source thesaurus management software, TemaTres[4], to manage and publish the pilot thesaurus online.

[3] http://apps.dri.ie/motif
[4] http://www.vocabularyserver.com/

Selecting terms for the Thesaurus I

The pilot thesaurus limited its scope to collecting approximately 350 terms from two chapters of Seán Ó Súilleabháin's *Handbook of Irish folklore* (Ó Súilleabháin 1942), and other vocabulary resources outlined below. These two chapters, on livelihood and household support, and nature, were chosen as they contained a broad range of terms, which could be used to demonstrate most thesaural relationships and facet analysis. It was, however, not possible to select all relevant terms from these chapters within the time frame of the project.

Additional terms were sourced from some article titles from *Béaloideas*, the journal of the Folklore of Ireland Society (Folklore of Ireland Society 2012), and Alan Dundes *The study of folklore* (Dundes 1965) as it provided a list of many forms of folklore. Most Irish terms were excluded from the pilot thesaurus and will need to be considered at a later date as part of a multilingual thesaurus. The exception to this was the term 'shebeen', which was included as an example of an Irish loan term.

Selecting terms for the Thesaurus II

During the selection process, terms for the pilot thesaurus were recorded in Microsoft Excel. Any associative relationships were also recorded. Definitions were included where any terms were ambiguous and qualifiers were used where necessary.

When considering the form of entry to be entered into the Excel spreadsheet, the pilot thesaurus followed the guidelines for English-language terms as set out by ISO 25964-1 as follows:

- Nouns and noun phrases — count nouns (dogs, cats) appear in the plural in the thesaurus. Non-count nouns (livestock) appear in the singular.
- Verbs — verbs take the gerund or verbal noun forms (fishing, hunting). The pilot thesaurus does not contain verbs in the infinitive.

- Adjectives — adjectives were avoided unless deemed significant to the subject.
- Adverbs — as per standard practice, adverbs were avoided in the pilot thesaurus.
- Articles — articles (a, the) were avoided unless they were an integral part of the term. If articles are used, equivalence relationships should be set up between the preferred term (which uses the article) and the non-preferred term (which does not use the article).

Terms with similar meaning were then grouped and preferred terms were chosen. Context was important to consider when choosing the terms for the thesaurus. For instance, non-scientific terms were chosen above scientific terms, where they appeared, as this matched the vocabulary found in the majority of the vocabulary resources. A similar style to the folklore literature (Selection I) was followed where possible.

In line with trends in thesaurus construction and best practice, and on the recommendation of international standards, the pilot employed the method of 'splitting' compound terms where practical and in such a way as to avoid inconsistency that is no more than two split components per compound concept. In practice, the terms assembled were comparatively simple and the issue of splitting compound concepts did not arise frequently.

Determination of the systematic structure

Facet analysis, with facets as top concepts, or top terms (TTs) in the hierarchies, was chosen as this is a more flexible structure, which can be more easily updated and the terms that were gathered logically fell into fundamental categories. This approach was also considered as a good demonstration of the hierarchy rules laid down by ISO 26594-1 that is that to form a proper hierarchy, narrower terms should fall into one of the following three relationship categories:

- The generic relationship where the broader and narrower terms form a genus/species or thing/kind relationship.

- The partitive relationship where they form a whole/part relationship.
- The instance relationship, which is used to name particular instances of a class of things.

At this stage, the decision was taken to structure the pilot thesaurus of Irish folklore using the fundamental facets as the main divisions while more detailed elements of the systematic structure were only made clear during the vocabulary analysis.

Vocabulary analysis using facet analysis

Terms in the vocabulary analysis process were recorded in Microsoft Excel for the initial analysis and Microsoft Word for the secondary analysis as this allowed for easier basic arranging and subdividing. In the first stage of the analysis, the selected terms were organised into basic groups. Following initial groupings, review and analysis, the fundamental facets decided on for the pilot thesaurus were:

- Time
- Place/space/environment
- Products
- Activities
- Processes and phenomena
- Events
- Agents
- Objects
- Materials
- Attributes and properties
- Parts

These fundamental facets were complemented by the facets of *Genre* and *Abstract entities and concepts*.

The next stage of the vocabulary analysis involved breaking down the facets into narrower divisions, using node labels to divide the facets into sub-facets or arrays and organising them according to the characteristics of division. The sub-facets and arrays should be considered as illustrative examples only and not as complete and exhaustive arrays as would be found in a complete thesaurus.

Addition of associative relationships, definitions, scope and other notes

Following the creation of the equivalence and hierarchical relationships, the terms were inputted into the thesaurus management software. Once all terms were present in the software, associative relationships were then added. This is the process recommended by *ISO 25964-1* as the most useful associative relationships are those that exist between hierarchies and therefore are easier to create when all hierarchies have been inputted into the thesaurus management software (ISO 2011).

As previously mentioned, the software chosen for the project was TemaTres. The advantages of using TemaTres were:

- It is open source software.
- All thesaural relationships are supported.
- It has out of the box functionality and a workable display.
- It uses export functionality in a number of formats, including SKOS-Core.

Disadvantages included:

- No comprehensive manual exists;
- Technical support was required for the initial set-up.

Definitions were added for a select number of terms and these definitions were taken from the Oxford English Dictionary Online. Scope notes were added to some of the top terms in the hierarchy to provide sample guidance on the placement of terms.

Print and electronic versions of the pilot thesaurus

Two lists, alphabetical and hierarchical, were then generated within TemaTres and exported as standard text files. These formed the basis of the print version of the thesaurus, which is now available online in both alphabetical and hierarchical forms.

Using the TemaTres software, an electronic version of the thesaurus now exists. It contains both hierarchical and alphabetical displays, which can be browsed as well as the ability to search by keyword. The electronic version is now available at http://apps.dri.ie/motif.

Review by experts

The guidelines were submitted for external review to experts in the field of cataloguing, metadata and knowledge organisation. The pilot thesaurus was submitted to subject experts for review and feedback, and we gratefully acknowledge the contributions of Dr Christoph Schmidt-Supprian (Trinity College Dublin); Jane Burns (Royal College of Surgeons in Ireland); Dr Jodi Schneider (INSIGHT, formerly DERI); Dr Elizabeth Mullins (University College Dublin); Críostóir Mac Cárthaigh (National Folklore Collection, UCD); and Anna Bale (National Folklore Collection, UCD).

Future work

Potential future work would include constructing a full thesaurus of Irish folklore by further engaging with experts in the folklore field and mapping to other vocabularies. It could also include the development of a multilingual thesaurus in both English and Irish.

Another logical extension of this project would be to engage with linked data professionals, represent the thesaurus in SKOS format and publish to the Semantic Web. By publishing folklore data in a structured, machine-readable format, the project will open up Irish data, which can then be exploited by web developers and computer scientists as well as information, heritage and research professionals. This dataset can then be linked to external data across the Web, adding additional context and value by linking together digitised collections of Irish interest.

Dissemination and engagement with librarians, archivists and other information professionals in the cultural heritage and enterprise communities will be an ongoing activity.

Bibliography

Dundes, A. (ed.) 1965 *The study of folklore.* London. Prentice Hall.

Folklore of Ireland Society. 2012 *Béaloideas*: the journal of the Folklore of Ireland Society.

International Organization for Standardization 2011 ISO 25964-1:2011, Information and Documentation. Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Geneva: International Organization for Standardization.

Ó Súilleabháin, S. 1942 *A handbook of Irish folklore.* Dublin. Educational Company of Ireland Ltd. for the Folklore of Ireland Society.

O'Carroll, A., and Webb, S. 2012 *Digital archiving in Ireland: national survey of the humanities and social sciences.* Maynooth. National University of Ireland.

Stewart, D. L. 2011 *Building enterprise taxonomies: a controlled vocabulary primer,* 2nd edn. Lexington, KY. Mokita Press.