Satu Niininen, Susanna Nykyri, Osma Suominen, (2017) "The future of metadata: open, linked, and multilingual – the YSO case", Journal of Documentation, Vol. 73 Issue: 3, pp.451-465, doi: 10.1108/JD-06-2016-0084

Purpose This paper focuses on the process of multilingual concept scheme construction and the challenges involved. The paper addresses concrete challenges faced in the construction process and especially those related to equivalence between terms and concepts. The paper also briefly outlines the translation strategies developed during the process of concept scheme construction.

Design/methodology/approach The analysis is based on experience acquired during the establishment of the Finnish thesaurus and ontology service Finto as well as the trilingual General Finnish Ontology YSO, both of which are being maintained and further developed at the National Library of Finland.

Findings Although URIs can be considered language-independent, they do not render concept schemes and their construction free of language-related challenges. The fundamental issue with all the challenges faced is how to maintain consistency and predictability when the nature of language requires each concept to be treated individually. The key to such challenges is to recognise the function of the vocabulary and the needs of its intended users.

Social implications Open science increases the transparency of not only research products, but also metadata tools. Gaining a deeper understanding of the challenges involved in their construction is important for a great variety of users – e.g., indexers, vocabulary builders and information seekers. Today, multilingualism is an essential aspect at both the national and international information society level.

Originality/value This paper draws on the practical challenges faced in concept scheme construction in a trilingual environment, with a focus on “concept scheme” as a translation and mapping unit.

Keywords controlled vocabularies, discourses, concept schemes, metadata, translation, equivalence, multilingualism

Paper type Case study

Satu Niininen, Susanna Nykyri & Osma Suominen

The future of metadata: open, linked and multilingual – The YSO case

1 Introduction

This paper focuses on the process of multilingual concept scheme construction and the challenges it involves. The analysis is based on experience gathered during the establishment of the Finnish thesaurus and ontology service Finto as well as the trilingual General Finnish Ontology YSO, both of which are being maintained and further developed at the National Library of Finland.

In this paper, *concept scheme* is understood as a vocabulary tool used for indexing and information retrieval in the same sense as thesauri and other types of vocabularies as described in ISO standards (see ISO 25964-1:2011, 4; 14). Furthermore, it has been understood as a tool to promote open access to information and science (Nykyri & Niininen 2015). We shall also discuss concept scheme as a translation and mapping unit, and briefly outline the translation strategies developed during the process of concept scheme construction.

With an overwhelming amount of data being published globally in a wide range of formats, locating and accessing relevant information is a challenge. In addition to better tools, there is also a more fundamental need for new working perspectives and practices, as firstly, the focus has shifted from records to entities, and secondly, the spectrum of users, needs and formats has expanded. The trend is to move from context-specific vocabulary tools towards collection-independent knowledge organisation systems (Zeng and Chan 2004), and thus the same vocabulary tool may nowadays be designed to serve the needs of the whole public sector, including science, administration and media. Furthermore, the context of the Semantic Web and linked open data has made it even more imperative to publish data in a format that is linkable and enriched with high-quality metadata.

It is also worth noting that a large amount of published material is not publicly available, and neither can it be openly accessible in the future. In many cases, however, the document

provider could still make the metadata freely available. Furthermore, in fields such as bibliometrics and digital humanities, metadata is treated as important research material in its own right, and research can suffer from gaps and errors in collection description and metadata management.

Figure 1: Isolated indexing - isolated data

Language plays a key role in participating in the global community. Through multilingual and open linked metadata, information can be located and retrieved not only across different collection providers, but also across languages. Consequently, resources indexed using one language can be retrieved using another, facilitating better access to information resources not published in one's native language. It should also be noted that non-English-speaking countries also publish a large amount of information resources in English and other world languages, but if such resources are indexed in their respective national bibliographies using a controlled vocabulary only available in the country's native language, much of this information can remain unseen to potential information seekers. In other words, the semantic dimension of linked open data can bring together resources across linguistic and organisational barriers.

As the majority of current solutions for accessing multilingual information are inadequate to answer the needs of increasingly diverse user groups from different cultural and linguistic backgrounds, it has become evident that traditional thesauri need to undergo a transformation to meet the new multilingual demands (Jorna & Davies 2001). Seeking and retrieving information across national borders is continuously on the rise, but the success of creating and using global information resources still depends on establishing a shared understanding of the concepts used (Nykyri 2010). Today, a change can be seen in practices, as new kinds of tools are being actively developed and the focus has increasingly shifted towards the challenges of different discourses and languages. Still, many central problems related to language are overlooked or handled from an unrealistic or unfruitful viewpoint (Hirst 2014).

2 Background

Why do we need controlled vocabulary tools in the free World Wide Web, particularly in the context of open science? Why is all the effort and guidance necessary? The obvious answer is because we operate with language. As Blair (2006, 2-3) summarises:

> "Information searches themselves inevitably require the searcher to ask for or describe the information he or she wants and to match those descriptions with the descriptions of the information that is available: in short, when we ask for or describe information we must mean something by these statements." (Ibid)

The inevitable problem is that the use of language involves a variety of perspectives and contexts. Wittgenstein sees language as a labyrinth of paths: "You approach from one side and know your way about; you approach the same place from another side and no longer know your way about" (Cited here Blair 2006, 28-29). To find our way in the labyrinth of language we need to understand how the different perspectives and contexts are shaped into *discourses*.

At a general level, *discourse* can be defined as the use of language in a social context (Pälli 2003, 22). The way each individual expresses their ideas and thoughts in a certain social environment is inevitably subjective, and in indexing, this can cause problems. In an information retrieval situation the selected search strategy and the synthetic structure of the search, as well as the language choices employed by the author, the indexer and the information seeker, all represent a type of discourse (Buckland, 1999; Nykyri 2010). It has been shown (Nykyri 2010) that in content management, discourse barriers may lead to greater differences than language barriers. Therefore greater differences may exist e.g. between Finnish indexers and Finnish social scientists than between Finnish and British indexers. There are also well-known challenges such as inflexibility (language in vocabularies being slow to reflect changes in natural language use), and documentary language may seem artificial or foreign in comparison to natural language (see e.g. Järvelin 1995; Cleveland & Cleveland 2001).

Information retrieval benefits from the use of controlled vocabularies in a number of ways. They not only solve problems related to (near-)synonymy but also allow the search to be broadened or narrowed according to their hierarchical structure (see e.g. Järvelin 1995, 180-184). However, more modern vocabulary tools such as concept schemes and ontologies are better equipped to acknowledge different discourses by using concepts identified by URIs (*Uniform Resource Identifiers*) instead of terms. By using URIs, a user can attain a variety of linguistic expressions for a single concept, as a single URI can represent different language terms of the concept as well as a number of alternative or variant terms. However, although using URIs makes referring to concepts easier, it does not eliminate the problematic nature of natural languages in vocabulary construction. As Hirst (2014, 7, 12) points out, we should have realistic and constructive expectations regarding a multilingual Semantic Web: "A Multilingual Semantic Web cannot rely on only an ontology as an interlingual representation or as a nonlinguistic representation for inference; there is, in practice, no clean separation between the conceptual and the linguistic", and that "the future of semantic representations for the Multilingual Semantic Web is likely to lie in imperfect nonsymbolic methods that work well enough in practice for most situations".

Although the article focuses on issues related to equivalence and the guidance concerning it, these are certainly not the only challenges. For instance, the corpus of a concept scheme should be carefully considered and designed. Traditionally, the corpus of controlled

vocabularies, such as the General Finnish Thesaurus YSA, is often based on the terms encountered in the indexed documents. This may lead to a biased or overly narrow understanding of an individual concept, which later may bring about unwanted search results in information retrieval. Furthermore, the increasingly heterogeneous users and audiences can be a double-edged sword: having a great variety of users can potentially lead to great benefits, but if the foundations are of poor quality or inadequate design the negative consequences can be extensive.

2.1 The *Finto Service and General Finnish Ontology YSO*

In a country such as Finland, operating in a trilingual context is nothing out of the ordinary. Finnish and Swedish are both national languages [1] and participating globally requires the command and use of English. The multilingual environment inevitably has its challenges, and managing information in such a setting requires the development of shared tools and practices.

The Finto service has been created to answer such needs with a tool that is also fully capable of operating in the context of linked open data and the Semantic Web. Finto is a Finnish service which enables the publication and browsing of thesauri, vocabularies and other concept schemes. The service also offers interfaces for integrating the concept schemes into other applications and systems (see finto.fi).

The service is being developed as a joint venture between the National Library of Finland, the Ministry of Finance, and the Ministry of Education and Culture. The project is a continuation of the work begun by the joint research project FinnONTO between Aalto University, the University of Helsinki and a consortium of over 30 other organisations from 2003 to 2012 (see Hyvönen et al. 2008; Lappalainen, Frosterus, Nykyri 2014; Suominen et al. 2014). The service is based on Skosmos, an open source publishing tool for SKOS vocabularies (for more information see skosmos.org).

Finto includes both general and domain-specific vocabularies and concept schemes as well as KOKO, which is a collection of interlinked Finnish concept schemes. KOKO is based on the General Finnish Ontology YSO which has been further refined and extended with domain-specific concept schemes such as the Ontology for Museum Domain, the Ontology of Applied Arts, and the Finnish Ontology of Photography. In this article the main focus is on the YSO.

The YSO is based on two separate thesauri, the General Finnish Thesaurus YSA and its Swedish version, Allärs. The thesauri were originally developed to be used primarily for the term-based indexing of printed library materials. Having been developed and maintained largely from the perspective of traditional library needs and practices, they are not suitable for use in the linked open data environment where it is necessary to link together actors and discourses across different databases and organisations. In practice, the transition from a traditional term-oriented thesaurus to a concept scheme has meant two major changes: firstly, the emphasis has moved beyond the term level to the concept level, and, secondly, the hierarchical structure

has been made complete and consistent so it can be used to broaden or narrow a search when necessary (see more in Lappalainen, Frosterus, Nykyri 2014).

The YSA and Allärs are monolingual; the YSO is trilingual covering Finnish, Swedish and English variants, and includes mappings to other vocabularies. Based on Finnish indexing needs, the YSO includes concepts of Finnish origins, e.g. *puukkojunkkarit* (a term used for troublemakers active in the Southern Ostrobothnia region of Finland in the 19th century), *Törnävä Church (*name of a church in the city on Seinäjoki*);* concepts foreign to Finnish culture, e.g. *samurais;* and concepts that are rather international, such as *symmetry*. In addition to more general concepts, the YSO also contains a significant number of concepts from various specific domains and is thus very applicable to indexing materials that are interdisciplinary and of varied themes.

As the term 'ontology' has sometimes been used quite liberally and even inconsistently to describe a variety of differing semantic resources, it can be difficult to gain a clear view of what actually makes an ontology. As Grabar et al. suggest, it may be more fruitful to treat ontologies and other semantic resources not as distinct types but as parts of the same continuum (2012, 375-376). The YSO and many other concept schemes included in Finto were originally created in the FinnONTO research project, and in that context were conceived as ontologies to both reflect their original representation language (Web Ontology Language OWL) and distinguish them from the monolingual term-based thesauri which they aim to replace. The creation of the YSO and the other FinnONTO ontologies was also influenced by similar work done elsewhere. When the FinnONTO project began in 2003, there was a wider trend of constructing ontologies based on thesauri and other types of controlled vocabularies (see e.g. Guarino & Welty 2002; Gangemi et al. 2002; Soergel et al. 2004).

As the YSO has indeed been built upon a thesaurus and is applied in broadly similar information indexing and retrieval use cases, it may also be seen as an advanced multilingual thesaurus. Since the YSO was originally conceived as an ontology, there has been a gradual shift from an OWL ontology towards a SKOS concept scheme. When the Finto service was launched, the representation language of the YSO's published version was changed from OWL to SKOS, with some extensions from ISO 25964. However, the name YSO has been retained both for historical reasons and to reflect the design of the concept hierarchy, which is based on principles of ontology construction.

Today, the YSO includes almost 30 000 concepts and it is designed to be used in the entire public sector. Its origin reflects document- and content-oriented indexing (see Fidel 1994 and Mai 2000), but today, a request-centred approach for maintaining and extending the vocabulary is being considered as well. In practice, concepts and the terms referring to them in the YSO need to be looked at from several perspectives in order to acknowledge the various different discourses involved. This can be done by providing scope notes, alternative labels and/or several broader terms, among other things.

With its roots firmly in a Finnish view of the world, the YSO is clearly not the easiest possible vocabulary to translate. Indeed, if aiming at the easiest possible translation corpus, its contents should not be constructed to serve a specific culture but rather should adopt a more general, international approach (see more in Nykyri 2010). However, in all multilingual communication, challenges are always inevitable and the Finnish context of the YSO does not as such make its basis *wrong,* but certainly more labour-intensive to manage. Moreover, it is important to remember that although playing a crucial role in content management, the YSO is still only one factor in a successful information retrieval situation, one which is affected by indexing guidelines and practices and the implementation of the information retrieval system.

2.2 *Standards and guidelines – How guided are we?*

Although the language-tagging facilities of the RDF data model provide the basic means for expressing multilingual lexical information, the multilingual challenges pertaining to linked data have been infrequently studied, and few recommendations are given on how to publish linked data in one or several languages (Gracia et al. 2012; Vila-Suero et al. 2014). This section presents an overview of how the construction of multilingual concept scheme has been guided by ISO standards which are broadly recognised and accepted as the leading authority for the field. The primary focus is on how the standards treat the concept of equivalence, and what practices are recommended for achieving an acceptable level of equivalence.

The principles for the construction of the YSO are based on the ISO standard *Information and documentation - Thesauri and interoperability with other vocabularies*, which consists of two parts, ISO 25964-1:2011 and ISO 25964-2:2013. The role of ISO 25964-1:2011 is to provide recommendations for the development and maintenance of thesauri intended for information retrieval applications, and it is applicable to both monolingual and multilingual thesauri (v-vi). ISO 25964-2:2013 focuses on describing, comparing and contrasting the elements and features of these vocabularies that are implicated when interoperability is needed, as well as giving recommendations for the establishment and maintenance of mappings between multiple vocabularies (v-vi).

The ISO standard recognises three approaches to the construction of multilingual thesauri:

1. Translation of a monolingual thesaurus
2. Merging of several distinct monolingual thesauri
3. Simultaneous construction of the various language versions of a multilingual thesaurus (ISO 25964-1:2011, 92)

In this regard, the YSO employs a hybrid approach. The Finnish and Swedish content is a result of merging the General Finnish Thesaurus and its Swedish version, Allärs. The resulting bilingual concept scheme has then been translated into English. In Dachelet's classification of multilingual thesaurus types (cited by Doerr, 2001), the Finnish and Swedish content in the YSO forms an *interlingua,* i.e. "a thesaurus made out of concepts that are created by fusing each cluster of similar concepts from different social groups into a new concept", while the English

content is a *translation*. This naturally has implications concerning the status of different language versions. The standard states that in a multilingual thesaurus, all languages should have equal status (ISO 25964-1:2011, 50). However, in the YSO, the underlying hierarchical structure is a reconciliation between the existing Finnish and Swedish thesauri, and thus the recommendation of equal status can only apply to the Finnish and Swedish terms. English terms have a secondary status as they do not comprise the foundation of the hierarchical structure. Section 3 provides practical examples of the implications these structural differences have on the translation process.

The standard widely recognises the elusive nature of full equivalence between languages and the limitations that natural languages can impose on the construction of controlled vocabularies (Ibid, 16, 50-57). The most frequently encountered levels of equivalence are described not as distinct relationship types but as “points along the spectrum of possibilities that lie between the extremes of exact equivalence and absence of equivalence” (Ibid, 51) [2].

The standard also offers several examples of typical problems and their suggested solutions, including instructions on how to manage issues with quasi-synonyms and homographs, the absence of an acceptable equivalent, and combined problems including several problematic aspects in one concept. Further examples of how these instructions are applied to the YSO are given in section 3.2.

With regard to mappings and interoperability, the standard recognises that vocabularies typically include different selections of concepts, and develop them into different levels of specificity which results in various equivalence situations requiring varied solutions (ISO 25964-2:2013, 21). Ideally, the source and the target vocabularies contain two identical concepts which can be mapped together with a simple one-to-one equivalence relation. In practice, however, the only available equivalence mappings can be hierarchical or associative in nature, or only available via compound equivalence mappings [3]. The practice of how these guidelines have been applied to mappings between the YSO and Library of Congress Subject Headings (LCSH) is further discussed in section 3.2.3.

With regard to the practical side of concept scheme construction, the standards are a valuable tool for analysing the different degrees of equivalence between languages and between vocabularies. However, as no standard can ever provide an exhaustive answer to all of the challenges faced, the construction process will always involve certain compromises. In the construction of the YSO, a key aim has been to follow the standards whenever applicable and to avoid solutions which would clearly go against the standards or reduce the precision of the translations or the mappings.

In addition to content-related guidelines, there is also a need for broader operating principles that govern the overall practices and perspectives of concept scheme construction. The guiding principle in our project has been that in order to be of use in the multilingual Semantic Web, the tools and practices in metadata production should fulfil the following criteria:

- The tools, methods and work practices are open and transparent

- The aim is shared between different actors
- The design is based on actual user needs and reflects the variety of user groups; different perspectives and needs are recognised and acknowledged
- Concepts should be shared, but their labels (i.e. terms) may reflect different kinds of discourses
- The use of uniform resource identifiers (URIs) allows data integration across languages, discourses and data providers (e.g. libraries, researchers)

3 The YSO process: From monolingual terms to multilingual concepts

This section discusses the practice of multilingual concept scheme construction with the primary focus being on the challenges of defining, reaching and maintaining a sufficient level of equivalence in cross-language communication. The section begins with a brief overview of the maintenance process.

3.1 The *maintenance process in brief*

The process of adding a new concept to the YSO begins with **analysis**, where the need for a new concept is determined. This usually occurs when the topic of a publication cannot be indexed with the existing concepts and the need arises to include a new concept. The concepts of the YSO thus reflect the language of science and literature, and the concept scheme is continuously updated with new topics. The following step involves naming the concept, i.e. giving it **preferred terms in Finnish and Swedish.** The concept is then placed in the **hierarchy** and is given associative relations to other concepts in order to provide rich contextual information for the indexer and the information retriever. If the Finnish and Swedish labels are mismatched in a way that makes it impossible to find a location that would accommodate the labels of both languages, a compromise must be made. This may involve selecting a location that is less than optimal for either of the languages, adjusting the preferred label of either language or further specifying the scope of the concept.

The concept is then given a **preferred term in English**, making contents indexed with the YSO visible to an international audience. However, it should be noted that at this point the concept has already been placed in the hierarchy, and the location will not be changed even if it is not ideal for the preferred English label. The next step is **mapping,** where the concept is linked to LCSH with the SKOS closeMatch [4] property if an applicable match is available. If the LCSH does not contain a suitable match, the YSO concept in question is left without a mapping, which is the case for approximately 57% of all YSO concepts. Initially, the LCSH was selected as the first mapping targets for the YSO because they are built for the annotation of library materials, as opposed to vocabularies such as Wordnet and Wikidata which are from different backgrounds. Furthermore, the LCSH forms a hub to which many library-oriented controlled vocabularies have already been linked, and could thus provide access across the library field.

Finally, the concept is given a URI and is **published in Finto** where it is available as linked open data and is open for integration and re-use.

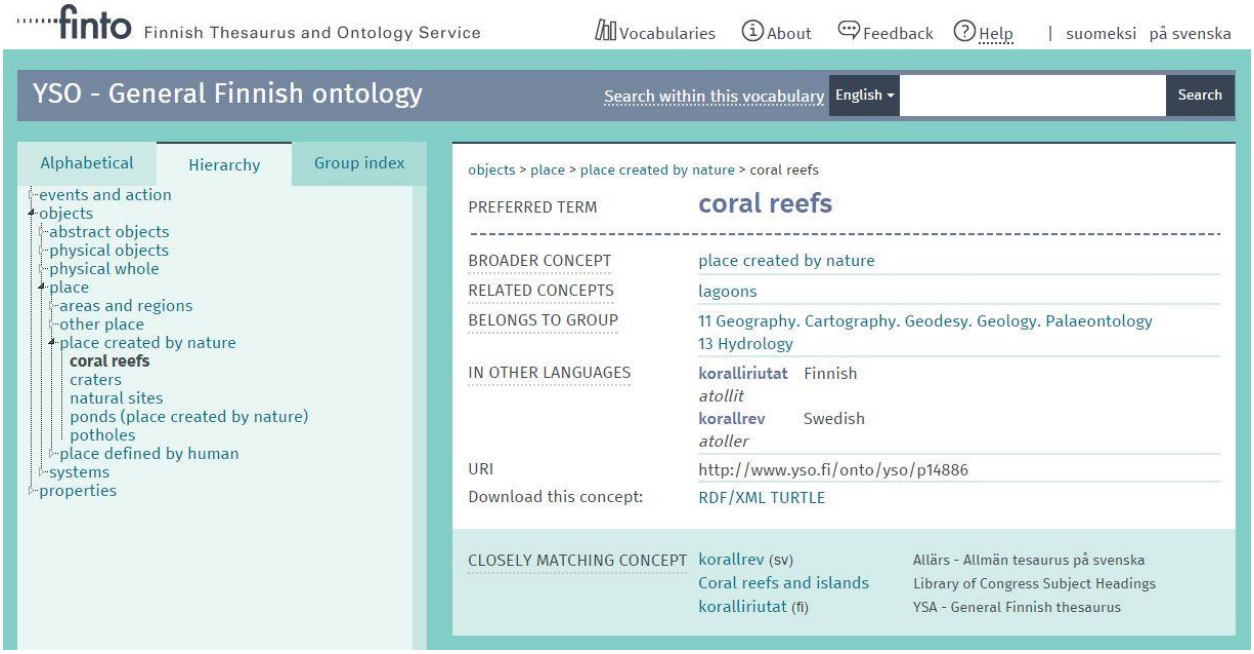

Figure 2: Coral reefs in the General Finnish Ontology YSO

3.2 *Challenges faced*

Each culture conceptualises the world from its own viewpoint, so meanings are seldom symmetrical across languages. Therefore, the aim has not been to pursue exact equivalence between languages but to instead lead the information retriever towards relevant search results regardless of which language is used in the query. However, a trilingual environment poses a number of language- and culture-related challenges, and building a complete and consistent hierarchy in more than one language is a complex process that requires compromises.

Challenges can be identified on multiple levels. In this article they have been loosely categorised as challenges between languages, challenges between cultures and challenges between vocabularies, in order to demonstrate typical and/or recurring problem types, and to present possible solutions. However, it should be noted that this categorisation is a loose framework only and cannot be regarded as an exhaustive representation of the challenges faced.

3.2.1 *Challenges between languages*

Challenges between languages refer to cases where a concept cannot be represented in the target language with a simple one-to-one equivalent. Typical examples of such are situations

where a concept is expressed using a linguistic or grammatical category not available in the target language, or situations where an accurate and symmetrical translation equivalent does not exist in the target language.

Various strategies are available for such situations. A typical solution is to use an explanatory qualifier, as in the concept of *decrease* in examples 1.1 and 1.2 below. Certainly a concept scheme constructed from an English-language or international viewpoint would most likely not include a distinction between active and passive decrease. However, in Finnish, the distinction is entirely relevant with *decrease (passive)* referring to instances where something decreases by itself without an external causing agent, and *decrease (active)* to instances where someone or something actively decreases something. A similar explanatory strategy has been applied to several concepts which are based on nouns derived from adjectives, as in example 1.2 below. The noun *eurooppalaisuus* is derived from the adjective *eurooppalainen,* meaning *European*. Thus *eurooppalaisuus* literally means the state of being European, and the translation is a compromise meant to convey the underlying meaning from one language to another. The translation strategy in all three examples is explanatory in nature and allows the reaching of an inexact level of equivalence in a situation where languages conceptualise phenomena in different ways. The procedure of accepting inexact or partial equivalents is recognised in ISO25964-1 as an acceptable solution (9.1, 9.3.1).

Example 1.4 is problematic due to differences in the context of use. The term *ennusteet* can be quite accurately translated into English as either *forecasts*, *predictions* or *prognoses*, depending on the context. However, the structure of the concept scheme allows only one preferred label, so a choice was made to select *forecasts* as the preferred label and *predictions* and *prognoses* as having a more limited scope of use as the alternative labels. Another way to deal with multiple target language equivalents can be seen in example 1.5. The Swedish language has three separate terms referring to different types of rivers with no specific equivalents in Finnish or English.

When exploring the YSO, it is important to keep in mind that although the concepts and the underlying phenomena they relate to are often global, they have been nevertheless conceptualised from the specific perspective of the Finnish and Swedish languages. In the translation process this can result in terms that mismatch with the concept's location in the hierarchy, as illustrated in example 1.6. It may seem odd for the English-speaking user to place *ant hills* in the hierarchy under *nests,* but in Finnish the hierarchy is entirely logical as *muurahaispesät* literally translates as *ant nests*. This obviously has its implications concerning the browsability of the hierarchy in English, but cannot be completely avoided as the English term is not considered during the hierarchy construction process.

	FI	**SV**	**EN**
1.1	väheneminen	minskning (antal)	decrease (passive)

1.2	vähentäminen	minskning (aktiv reducering av antal)	decrease (active)
1.3	eurooppalaisuus	europeisk identitet	European identity
1.4	ennusteet	prognoses	forecasts alt: predictions alt: prognosis
1.5	joet	floder, åar och älvar	rivers
1.6	pesät ↳ muurahaispesät (literally ant nests)	bon ↳ myrstackar	nests ↳ ant hills

Example set 1: Challenges between languages

3.2.2 *Challenges between cultures*

All cultures are founded on their own sets of systems, values and practices, so their languages carry different understandings of the world. This section highlights issues pertaining to concepts that are native to a certain cultural sphere and do not translate very well into another. Here the category of culture-specific concepts is understood as a broad continuum ranging across different levels of specificity. At one end are concepts referring to phenomena only found in a specific culture which often lack translation equivalents altogether, as the concepts do not exist in the target culture. Such concepts include names of societal structures, traditions and cultural phenomena limited to a specific culture, and e.g. names of professions and vocations which often reflect the underlying societal structure. At the other end of the spectrum are concepts with more subtle and implied manifestations of cultural differences. In these cases the denotational meaning is often simple enough to translate but the concepts can carry overtones that are challenging to convey, or their use may be limited to specific contexts that the translation equivalent cannot express.

When a translation equivalent is completely unavailable, the solution can be to borrow the term used in the source language and use it as a citation loan in the target language. However, such loans should always be accompanied by a clarifying scope note, offering the English-speaking user a brief explanation of the concept and its use. Furthermore, this strategy is mostly employed for terms referring to traditional or historical concepts, as seen in example 2.1 in example set 2 below. For contemporary concepts that lack an equivalent, a more preferred strategy has been to use a paraphrase or partial translation which conveys the most essential characteristics of the concept and can be complemented with an alternative label, a scope note and/or qualifier if necessary (see examples 2.2 and 2.3).

When the concept carries overtones that are missing from the closest possible target language equivalent, it is often advisable to select a descriptive strategy and add a qualifier in parentheses or a clarifying scope note in order to prevent misunderstandings, as illustrated in example 2.4.

If the scope of a concept is more specific than what the translation implies, or the concept is meant to be used in a particular context only, the translation must be complemented with a scope note and/or qualifier, as seen in examples 2.5 and 2.6. There the distinction between *perquisites* and *employee benefits* is made according to Finnish tax legislation which the international user cannot be expected to be familiar with. When the scope notes refers to other concepts a reciprocal note should be added to all the concepts mentioned (ISO 25964-1:2011, 21).

	FI	**SV**	**EN**	**Scope note**
2.1	helavyöt	söljebälten	helavyöt	Particular type of belt used with Southern Ostrobothnian national costumes.
2.2	liikuntalukiot	idrottsgymnasier	general upper secondary schools focusing on sport and exercise	
2.3	ryhmäkasvit	gruppväxter	plants used in groups	
2.4	kirkonkirjat	kyrkböcker	church registers	Only used in historical contexts, otherwise use population registers.
2.5	luontoisedut	naturaförmåner	perquisites	Refers to taxable benefits. For non-taxable benefits, see employee benefits
2.6	henkilökuntaedut	personalförmåner	employee benefits	Refers to non-taxable benefits. For taxable benefits, see perquisites.

Example set 2: Challenges between cultures

3.2.3 *Challenges between vocabularies*

Translating a complete concept scheme into English and linking the concepts to the LCSH when applicable equivalents are available involves connecting the languages of two very different cultural spheres. Although both vocabularies are used for indexing library materials, they are built upon their own constructions, practices and history. Currently, the YSO comprises nearly 30,000 concepts, of which approximately 43% can be linked to LCSH concepts, which total nearly 340,000 [5].

In terms of their hierarchical structure, the YSO and LCSH are not symmetrical. Unlike the YSO, the LCSH does not employ a complete hierarchy. Instead, the LCSH consists of several smaller and non-connected hierarchies resulting in a large number of top-level terms. Due to structural as well as cultural differences between the YSO and LCSH, we have not attempted to perform what Doerr (2001) calls a *complete* mapping, which would require assigning every YSO concept either an exact equivalence in the LCSH or at least one broader and one narrower equivalence (when available), a process which would have been very labour-intensive. For practical reasons, we have thus decided to map only those YSO concepts for which a suitable equivalent or near-equivalent concept exists in the LCSH.

The preferred labels need not be identical in order to establish linkage, but in such cases the context and scope of both labels must be checked. As illustrated in example 3.1 below, *corporate executives* is eligible for linkage with the LCSH concept *executives,* as the two are very likely to be used to index the same or similar materials. Moreover, differences in the level of specificity are not considered an impediment if the difference is regarded as minor, as in example 3.2 where *rubber boots* is linked to the LCSH concept *Rubber footwear*. However, defining a sufficient level of equivalence appears to be as much art as it is science in the sense that each problematic concept needs to be evaluated individually. As seen in example 3.3, the closest possible mapping equivalent for the YSO concept *retroactivity* is the LCSH concept *Retroactive laws*. Both concepts represent a different conceptualisation of the same legal phenomenon, but because they are expressed with such different formulations it would be potentially misleading to mark them as equivalents. Such cases of near-equivalence could be linked with the SKOS:related property, used to state associative mapping links between concepts [6]. The property is not currently used in the YSO but could be added later if it is deemed necessary to enrich the LCSH mappings with more thematic and contextual information.

Compound equivalence mappings are accepted in cases where a concept is represented in the target vocabulary by a combination of one or more concepts, i.e., cumulative equivalence as presented in section 2.2. As seen in examples 3.4 and 3.5 below, the YSO contains two separate concepts, *pilots (shipping)* and *pilotage (shipping),* whereas the LCSH has incorporated them into the single concept *pilots and pilotage*. In such cases both YSO concepts can be linked to the same LCSH concept. However, intersecting equivalence mappings as explained in section 2.2 are not accepted by the YSO as their information value has not been deemed sufficient.

The most typical case of mismatch between the YSO and LCSH is a complete lack of a feasible linking equivalent. Such cases include a high number of culture-specific concepts of Finnish culture, although due to the extensiveness of the LCSH, many inherently Finnish concepts are also included (e.g., *Finlandia Ski Race*). However, not all cases of non-equivalence are due to the culture-specific character of the concepts. In fact, most of the time when a mapping equivalence is not available it is due to the different structure of the vocabularies or a different conceptualisation of a shared phenomenon. A number of fairly general YSO concepts, such as *biological children, downshifting, synergy* and *barons* are currently not included in the LCSH.

The YSO and LCSH also differ in their position with respect to pre-coordinated indexing. In fact, the YSO's few pre-coordinated indexing strings are currently under revision and likely to be uncoupled altogether whereas the LCSH continues to employ pre-coordination. This is resulting in situations where essentially the same concept is expressed in one vocabulary with a compound phrase and in another with a pre-coordinated string, and a decision must be made whether or not to link them together. Mappings have been established if the concepts both cover a similar limited scope and could be used interchangeably, as in example 3.6. Furthermore, the LCSH employs a large number of subdivisions which are used in heading-subdivision combinations, whereas the YSO does not employ such a concept type or have such a practice. In order to still be able to utilise the numerous subdivisions, a compromise was made allowing for them to be mapped to YSO concepts, with the SKOS:narrower property indicating a hierarchical equivalence [7].

	FI	SV	EN	LCSH mapping
3.1	yritysjohtajat	företagsledare	corporate executives	Executives
3.2	kumisaappaat	gummistövlar	rubber boots	Rubber footwear
3.3	taannehtivuus	retroaktivitet	retroactivity	≠ Retroactive laws
3.4	luotsit	lotsar	pilots (shipping)	Pilots and pilotage
3.5	luotsaus	lotsning	pilotage (shipping)	Pilots and pilotage
3.6	haimataudit	pankreassjukdomar	pancreatic diseases	Pancreas-Diseases

Example set 3: Challenges between vocabularies

3.3 *Conclusions*

To sum up, the challenges faced are mostly equivalence-related, regarding the definition of either translation or mapping equivalence. In addition to defining a sufficient level of closeness that the equivalents should achieve, it is also crucial to remain consistent in the choices made. However, the mapping-related challenges tend to be slightly less problematic as a concept can be left without mapping if a suitable match does not exist, whereas a translation equivalent must always be established. Furthermore, the limited browsability of the English language version has been acknowledged as a disadvantage but with the current construction procedure there is no feasible way to avoid this.

A very frequent challenge is the difficulty of making distinctions between concepts that are not commonly used as separate concepts in either of the languages or that convey phenomena not existing in the target culture. However, these cases can often be resolved by resorting to inexact equivalence and/or incorporating a qualifier to narrow down the meaning of the concept in question. Therefore these cases, however frequent, do not pose a particularly demanding problem.

The most time-consuming though less frequent type of problem is the explication of ambiguous translation equivalents with limited space to explain the implied or context-related differences in the scope of their use. Such cases often require selecting a single translation equivalent as the preferred label for concepts that can be expressed through a range of terms depending on context. A qualifier is often needed, but as the ambiguity tends to be context-related (i.e. this concept can be used in *these* contexts but not *those*), it is challenging to find an exhaustive definition.

The fundamental issue with all the challenges faced is how to maintain consistency and predictability when the nature of language requires each concept to be treated individually. Based on the experience gathered in the Finto project, it seems that the key to such challenges is to recognise the function of the vocabulary and the needs of its intended users. In order to construct a vocabulary that serves its purpose as an indexing tool it is best to consider what implications a certain solution would have for information-seeking situations.

4 Discussion

In this article the emphasis has been on language and multilingualism in concept scheme construction. As shown here, although URIs can be considered language-independent, they do not free the construction process from language-related challenges. Translating or creating multilingual concept schemes and mapping them to other resources needs to be studied and discussed more carefully and in different contexts (see e.g. Doerr 2001; Zeng and Chan 2004; Trojahn, Quaresma, Vieira 2008 & 2010; Helou et al. 2016). It should also be noted that although automatic translation technology has improved considerably in recent years, the quality of automatic translation still varies immensely, and corrective human review generally gives the best results (Embley et al. 2011). Thus, technology can be very helpful, but human effort is still very much needed.

The great variety of user contexts means new kinds of challenges and practices. According to our experience, it is highly essential to establish close co-operation between the concept scheme developers and the adopters at an early stage of the development process and to ensure that the aims are well-justified and shared. In order to shift from a traditional thesaurus

founded on library collections to a machine-readable concept scheme designed to operate in the Semantic Web, the working processes must be enriched to cover an even greater variety of users and materials. Subject access is a powerful gateway to information, and by providing effective tools and developing shared practices we can ensure long-term accessibility to our dynamic and changing information environment.

However, long-term accessibility does not come without commitment to continuous maintenance. An essential question for further discussion is the time aspect: how to ensure that the vocabulary keeps up with language, which is in constant flux. Not only the construction, but also the maintenance of vocabulary tools is very labour-intensive and requires a permanent allocation of resources. Without upkeep the vocabulary can begin to diverge from actual language use and eventually become obsolete.

When evaluating the end-result, a shift should be made closer to the information seeker. It is important to remember that the final aim of controlled vocabularies – including multilingual concept schemes is to provide better search results and easier access to information. As Hirst (2014, 8) states:

> "The Semantic Web vision rightly emphasizes the benefit of the *information seeker*, whose task will be made easier and who will be given a greater chance of success. The benefit to the *information provider*, who wants to bring their information to the notice of the world for commercial, administrative, or other purposes, is secondary and often indirect." (Ibid)

Moreover, to be useful to information seekers, the design and content of the concept scheme should also be open and transparent and not designed to favour certain information providers as the commercial sector tends to now do. The long-term consequences of adopting closed systems can be unpredictable as well as unfavourable, and carry a risk of libraries becoming gatekeepers instead of information providers.

The challenge of constructing and harmonising multilingual metadata is a crucial element in the context of the global linked open data environment. However, this cannot be achieved without acknowledging the differences between the specific characteristics of different languages.

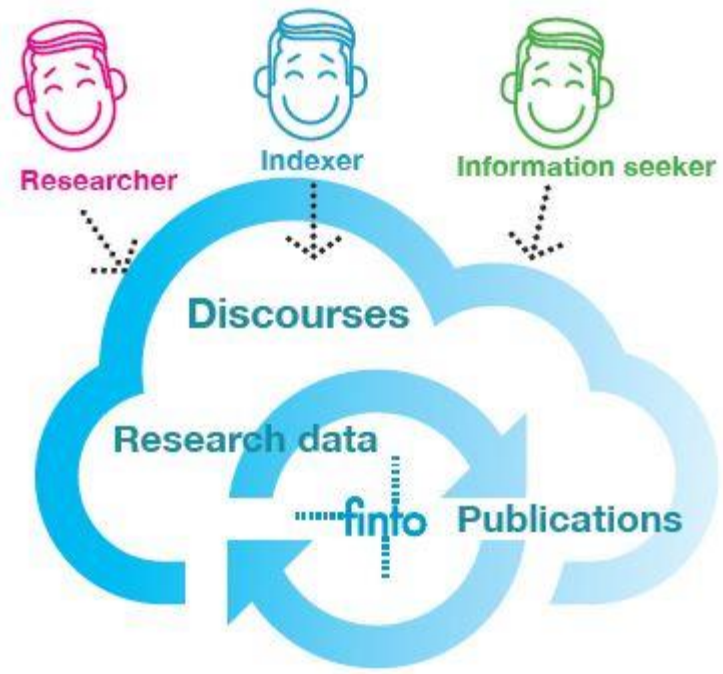

Figure 3: Shared tools and practices – shared data

5 Notes

[1] Finland Swedish is a dialect of Swedish spoken in Finland by the Swedish-speaking population as their mother tongue. Finnish and Swedish are the two national languages of Finland.

[2] The levels of equivalence are divided into the following categories:

a. **exact equivalence** where the preferred term is culturally and semantically equivalent in every language of the thesaurus
b. **inexact** or **near-equivalence** where the terms carry minor differences in scope due to differences in culture, connotation or appreciation but are still close enough to represent the same concept in the thesaurus
c. **partial equivalence** where a concept can be represented only through a term which is normally considered to represent a broader or narrower aspect of the concept but could be admissible into the thesaurus if the difference in scope is considered to be small enough.
d. **non-equivalence** where no term can be found to provide even a partial or inexact equivalence. (ISO 25964-1:2011, 51-52)

[3] The equivalence mapping types are as follows:

1. simple equivalence
2. compound equivalence (one to many equivalence), which covers the two distinct types of equivalence, namely
    2.1. intersecting as between *women executives* in one vocabulary and *women* + *executives* in another
    2.2. cumulative as between *hosiery* in one vocabulary and *stockings* + *socks* in another. (ISO 25964-2:2013, 21-24)

[4] A skos:closeMatch assertion indicates that two concepts are sufficiently similar that they can be used interchangeably in applications which consider the two concept schemes they belong to. However, skos:closeMatch is not defined as transitive, which prevents such similarity assessments from propagating beyond these two schemes. (Isaac et al.,. 2009)

[5] A brief introduction to the structure and use of the LCSH is available on the Library of Congress website. (Library of Congress Subject Headings PDF files, available at: https://www.loc.gov/aba/publications/FreeLCSH/freelcsh.html#About)

[6] The skos:related property enables the representation of associative (non-hierarchical) links, such as the relationship between one type of event and a category of entities which typically participate in it. Another use for skos:related is between two categories where neither is more general or more specific. (Isaac et al.,. 2009)

[7] The skos:narrower property is used to assert the inverse, namely when one concept is narrower in meaning (i.e. more specific) than another. (Isaac et al.,. 2009)

6 References

Blair, D.C. (2006), *Wittgenstein, language, and information : back to the rough ground!*, Springer, Dordrecht.

Buckland, M.K. (1999), "Vocabulary as a Central Concept in Library and Information Science", in Aparac, T., Saracevic, T., Ingwersen, P. and Vakkari, P. (Eds.) *CoLIS3 Proceedings: DIGITAL LIBRARIES: Interdisciplinary Concepts, Challenges and Opportunities*, Benja Publishing, Lokve.

Cleveland, D.B. and Cleveland, A.D. (2001), *Introduction to indexing and abstracting*, 3rd ed., Libraries Unlimited, Englewood, Colorado.

Doerr, M. (2006), "Semantic Problems of Thesaurus Mapping"*, Journal of Digital Information,* Vol. 1, No. 8.

Embley, D.W., Liddle, S.W., Lonsdale, D.W. and Tijerino, Y. (2011), "Multilingual Ontologies for Cross-Language Information Extraction and Semantic Search", in Jeusfeld, M., Delcambre, L. & Ling, (Eds.)*, Conceptual Modeling -- ER 2011: 30th International Conference, ER 2011, Brussels, Belgium, October 31 - November 3, 2011. Proceedings* Springer, Berlin, Heidelberg, pp. 147-160.

Fidel, R. (1994), "User-centered Indexing"*, Journal of the American Society for Information Science,* Vol. 45, No. 8, pp. 572-576.

Grabar, N., Hamon, T. and Bodenreider, O. (2012), "Ontologies and terminologies: Continuum or dichotomy?"*, Applied ontology,* Vol. 7, No. 4, pp. 375-386.

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P. and McCrae, J. (2012), "Challenges for the multilingual web of data"*, Web Semantics: Science, Services and Agents on the World Wide Web,* Vol. 11, pp. 63-71.

Helou, M.A., Palmonari, M. and Jarrar, M. (2016), "Effectiveness of Automatic Translations for Cross-Lingual Ontology Mapping"*, Journal of Artificial Intelligence Research,* Vol. 55, pp. 165-208.

Hirst, G. (2014), "Overcoming Linguistic Barriers to the Multilingual Semantic Web", in Buitelaar, P. & Cimiano, P. (Eds.)*, Towards the Multilingual Semantic Web: Principles, Methods and Applications* Springer, Berlin, Heidelberg, pp. 3-14.

Hyvönen, E., Viljanen, K., Tuominen, J. and Seppälä, K. (2008), "Building a national semantic web ontology and ontology service infrastructure - the FinnONTO approach", in Bechhofer, S.; Hauswirth, M.; Hoffmann J. & Koubarakis, M. (Eds.)*, The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings* Springer, Berlin Heidelberg, pp. 95-109, available at http://seco.cs.aalto.fi/publications/2008/hyvonen-et-al-building-2008.pdf (accessed 07 June 2016).

Information and documentation : Part 1, Thesauri and interoperability with other vocabularies (2011), International Organisation for Standardization, Geneva.

Information and documentation : Part 2, Thesauri and interoperability with other vocabularies (2013), International Organisation for Standardization, Geneva.

Isaac, A. and Summers, E. (2010), "SKOS Simple Knowledge Organization System Primer, 2008" available at: URL http://www.w3.org/TR/2009/NOTE-skos-primer-20090818 (accessed 15 June 2016).

Jorna, K. and Davies, S. (2001), "Multilingual thesauri for the modern world - no ideal solution?"*, Journal of Documentation,* Vol. 57, No. 2, pp. 284-295.

Järvelin, K. (1995), *Tekstitiedonhaku tietokannoista : johdatus periaatteisiin ja menetelmiin,* Suomen atk-kustannus, Espoo.

Lappalainen, M., Frosterus, M. and Nykyri, S. (2014), "Reuse of library thesaurus data as ontologies for the public sector", *paper presented at IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 86 - Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Programme. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France*, available at http://library.ifla.org/819/1/086-lappalainen-en.pdf (accessed 15 June 2016).

Mai, J. (2000), *The subject indexing process: an investigation of problems in knowledge representation*, Unpublished Doctoral Dissertation, University of Texas at Austin.

Nykyri, S. (2010), *Equivalence and translation strategies in multilingual thesaurus construction,* Åbo Akademi University Press, Åbo, available at http://urn.fi/URN:ISBN:978-951-765-521-7 (accessed 15 June 2016).

Nykyri, S. and Niininen, S. (2015), "The future of metadata: Open, linked and multilingual"*, Scandinavian Library Quarterly,* Vol. 48, No. 1-2, pp. 13-15. available at: http://slq.nu/?article=volume-48-no-1-2-2015-4 (accessed 15 June 2016).

Potter, J. (1990), "Discourse: Noun, verb or social practice?"*, Philosophical Psychology,* Vol. 3, No. 2, pp. 205-217. available at: https://www.researchgate.net/publication/240239892_Discourse_Noun_verb_or_social_practice (accessed 15 June 2016).

Pälli, P. (2003), *Ihmisryhmä diskurssissa ja diskurssina,* Tampere University Press, Tampere.

Suominen, O., Pessala, S., Tuominen, J., Lappalainen, M., Nykyri, S., Ylikotila, H., Frosterus, M. and Hyvönen, E. (2014), "Deploying National Ontology Services: From ONKI to Finto", *paper presented at the ISWC 2014 Industry track, 19-23 October 2014, Trentino, Italy*, available at: https://www.seco.tkk.fi/publications/2014/suominen-et-al-deploying-onki-finto-2014.pdf (accessed 15 June 2016).

Trojahn, C., Quaresma, P. and Vieira, R. (2008), "A Framework for Multilingual Ontology Mapping", *paper presented at the Sixth International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco,* available at:

http://www.di.uevora.pt/~pq/papers/lrec08.pdf *(accessed 15 Jun 2016)*.

Trojahn, C., Quaresma, P. and Vieira, R. (2010), "An API for Multi-lingual Ontology Matching", *paper presented at the Seventh International Conference on Language Resources and Evaluation, 17-23 May 2010, Valletta, Malta*, available at: http://www.lrec-conf.org/proceedings/lrec2010/pdf/691_Paper.pdf (accessed 15 June 2016).

Vila-Suero, D., Gómez-Pérez, A., Montiel-Ponsoda, E., Gracia, J. and Aguado-de-Cea, G. (2014), "Publishing Linked Data on the Web: The Multilingual Dimension", in Buitelaar, P. and Cimiano, P. (Eds.), *Towards the Multilingual Semantic Web,* Springer, Berlin Heidelberg, pp. 101-117.

Zeng, M.L. and Chan, L.M. (2004), "Trends and issues in establishing interoperability among knowledge organization systems"*, Journal of the American Society for Information Science and Technology,* Vol. 55, No. 5, pp. 377-395.

Acknowledgments: We thank Thomas Baker for his insightful comments about concept schemes and their relationship to thesauri and ontologies.