Towards a general Norwegian thesaurus?

Subproject *Methodology for mapping Humord to WebDewey*

by Are Dag Gulbrandsen, Dan Michael O. Heggø, Unni Knutsen, and Grete Seland

Presentation of the mapping project at the UiO Library, based on the pilot project report as per March 1st 2015

Translated and with some updates by Elise Conradi and Grete Seland

Introduction

Background

The one-year pilot project *Methodology for mapping Humord to WebDewey* at the University of Oslo Library (henceforth abbreviated UiO Library) was a sub-project of the larger project *Towards a general Norwegian thesaurus?* The results of the project were reported to the National Library of Norway in March. The project has received new funding and is prolonged as the two-year project *Mapping to Norwegian WebDewey*.

The present document provides a presentation of the mapping project at the UiO Library, at the point when we closed the pilot and continued our efforts in the prolonged project. The paper is largely a translation of the pilot project report as per March 1st 2015, enriched with some updates and comments. It is indeed an intellectual work in progress, so the discussions, opinions and solutions presented below are under constant debate and review in our project group. This "state of the art" description of our challenges in mapping a thesaurus to WebDewey is intended as a starting point for our joint discussions at the EDUG seminar in Naples.

Participants in the pilot project have been: Are Gulbrandsen, Dan Michael O. Heggø, Berit Sonja Hougaard, Viola Kuldvere, Vibeke Stockinger Lundetræ (from June 1st 2014), Mari Lundevall, Grete Seland (from January 15th 2015), and Unni Knutsen (project leader). They all continue their work in the present project *Mapping to Norwegian WebDewey*

The main objective of the pilot project has been to develop a methodology for mapping as a computer-assisted intellectual task. Our point of departure has been that this approach will provide a better basis for making correct mapping decisions than if the task is performed solely as a manual procedure. Moreover, it will require less human resources.

A short introduction to the vocabularies

Humord ('humanities terms') originated in 1993/1994 as a thesaurus for the humanities, and has since 2011 included terms from the social sciences. It is intended for post-coordinated indexing. Humord currently comprises approximately 18 500 main terms and about 8 500 references. Humord is from the outset developed as a collaborative project. The indexing partnership group currently consists of the university libraries in Oslo, Bergen and Tromsø, as well as the Norwegian Center for Studies of Holocaust and Religious Minorities. Development work at Humord is coordinated by a group which is chaired by the Humord coordinator at UiO Library.

Realfagstermer ('science terms') is a controlled, pre-coordinated vocabulary of subject headings that mainly covers the fields of science, mathematics and computer science. The vocabulary includes approximately 15 000 main terms and about 2 000 references.

Both vocabularies consist of terms in Norwegian. Realfagstermer also contains terms in other languages (e.g., English and Latin).

What is mapping, and why is mapping important?

Mapping can be defined as an activity in which relationships (i.e. mappings) between concepts in two different controlled vocabularies are established. The set of relationships between two vocabularies is called a crosswalk (i.e. a transition) between the vocabularies.

Crosswalks, either between two thesauri, or between a thesaurus and a classification scheme, allow for interoperability, e.g. in searching across vocabularies. Locally, a crosswalk between Humord and Dewey will open up for the possibility to apply Humord subject headings as search terms, and retrieve documents that are not indexed with Humord, but classified with Dewey.

Outline of the report

This report is arranged according to the following structure:

First, we will present the ISO standard 25964-2 for interoperability between vocabularies. This standard forms the theoretical and methodological foundation on which our work rests. We give an account for our choice of mapping model, present the different mapping relationships, and explain the overall choices we have made based on the recommendations of the standard and the nature vocabularies. We also comment upon the relationship between the ISO standard and SKOS[1].

The next chapter will discuss intellectual challenges associated with mapping – both general considerations on mapping, and more specific issues related to mapping between the vocabularies Humord and Dewey. We present the challenges we face in the selection of mapping candidates, as well as the considerations of relationship type for each mapping. This spring we are undertaking a number of test mappings, to uncover the complexities that will arise in connection with the actual mapping procedure.

As part of our goal of applying an approach with computer-assisted mapping, we will present the mapping software tool which is developed as part of the project. In the chapter on design of the mapping tool ccmapper, we will introduce the data sources which will be used as input for the algorithms which will produce mapping candidates. The SKOS-model, as well as various methods used in the tool (e.g., term weighting and lemmatization) will be presented.

In conclusion, we will draw some lines from the results of the pilot project and the work to be performed in the next project.

[1] Simple Knowledge Organization System http://www.w3.org/2004/02/skos/

ISO-25964-2 and SKOS as a point of departure for mapping

As mentioned in the introduction, the ISO-standard 25964, (International Organization for Standardization, 2009, 2013) constitutes the theoretical and methodological foundation for our mapping work. ISO 25964-1 primarily describes how a thesaurus is constructed and gives suggestions regarding best practice in this area. ISO 25964-2 *Interoperability with other vocabularies* stems from the recognition that there is a pressing need to identify and localize relevant information across large collections. This leads to the necessity of creating semantic interoperability. For retrieval purposes, the goal of interoperability between vocabularies is to connect a concept from one vocabulary to a corresponding concept in one or more other vocabularies. ISO 25964-2 offers basic descriptions of other types of vocabularies (like classification systems) and recommendations for mapping between these and thesauri. Of the two parts of the standard, it is therefore ISO 25964-2 that has had the greatest meaning for our work. References hereinafter to the ISO-standard are implicitly to ISO 25964-2.

ISO 25964-2 defines the verb *map* in the following way: "to establish relationships between the concepts of one vocabulary and those of another", while *mapping* is defined as "the process of establishing relationships between the concepts of one vocabulary and those of another" (International Organization for Standardization, 2013, p. 7).

Choice of structural model

In the introduction, the standard presents various structural models for how vocabularies can be mapped together. The standard emphasizes that the structure of the model and the direction of the mapping should be clarified in the beginning of all mapping projects. Good advice is given for the choice of structural model.

The natural choice for us was to pursue a hub-model (figure 1) in which the goal is to map our vocabularies (Humord and Realfagstermer) to the Dewey Decimal Classification system (the Norwegian translation). When Humord is mapped to DDC, and Library of Congress Subject Headings (LCSH) is mapped to the same target, we don't see the need to invest major resources to also map Humord directly to LCSH. In a future end-user system, we envision a user who, by searching for a term in Humord, will automatically be able to continue the search in databases and services with documents indexed with LCSH. The underlying connector will be a Dewey notation. The weakness of this methodology is that a Dewey notation often contains more than one subject. A connection based on mapping via a classification notation will therefore result in the retrieval of some irrelevant documents. Our assessment, however, is that since the DDC has a high degree of granularity and essentially collates related topics, this will not create a major problem for the end-user.

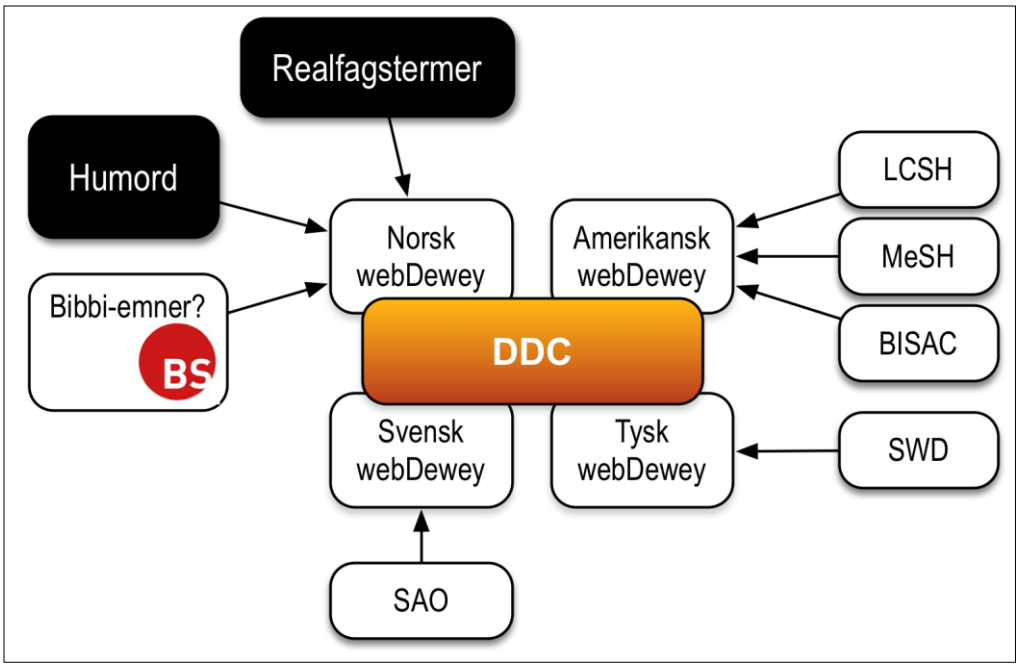

Figure 1: Dewey Decimal Classification (DDC) as a hub. Humord and Realfagstermer are here mapped to the Norwegian WebDewey. Other vocabularies are mapped to DDC in corresponding ways.

As shown in figure 1, the use of DDC as a hub is a widespread international strategy. The ISO-report recommends this approach for "the reconciliation of vocabularies that have been independently developed and/or have already been applied to collections" (s. 20).

When it comes to the direction of the mapping, the ISO-standard uses the following concepts: *source vocabulary* and *target vocabulary*. Source vocabulary is defined as "a vocabulary that serves as a starting point when seeking a corresponding term or concept in another vocabulary" (p. 12), while a target vocabulary is defined as "a vocabulary in which a term or concept is sought corresponding to an existing term or concept in a source vocabulary" (p. 13). In our project, Humord is the source vocabulary to be mapped to the target vocabulary DDC.

It is important to be cognizant of the fact that we are mapping the meaning of words. This means that we must be conscious of the fact that concurrent words do not necessarily have the same meaning.

Various types of mapping relationships

When two terms from distinct vocabularies are mapped together, one must establish what type of relationship exists between them. Mapping relationships are based on relationship types known from thesaurus construction, with equivalence, hierarchical and associative connections. The ISO-standard operates with two types of equivalence relationships. The hierarchical relationship type goes both ways (*broader* and *narrower*). This gives us the following five relationships:

=EQ	(EQ = Equivalence). The equal sign indicates that the mapping is exact
~EQ	The tilde symbol indicates that the mapping is not exact. This means that the concepts can be similar in some contexts, but not in all, or that the concepts can be partially overlapping or differ a bit in meaning
BM	Broader mapping. The term in the target vocabulary has a wider meaning than the term in the source vocabulary
NM	Narrower mapping. The term in the target vocabulary has a more specific meaning than the term in the source vocabulary
RM	Related mapping. The term in the target vocabulary is associated with the term in the source vocabulary, but is not a synonym, a quasi-synonym or a broader or narrower term

One type of equivalence relationship is *compound equivalence mapping* where a concept in one vocabulary is equivalent to several concepts in another vocabulary. One can then use a combination of mapping relationships and Boolean operators (AND and OR, symbolized with + and │). One of the examples in the ISO-standard (p. 36) shows how the thesaurus term "institutions" may be mapped to three relevant classification notations: "institutions EQ E100 │ H100 │ D100". However, this cannot be expressed in SKOS. Alternatively, one can use several independent mappings. We plan to use the latter solution, thereby avoiding the use of Boolean operators

Compound equivalence mapping may also be relevant in cases where one of the vocabularies contains compound terms while the other vocabulary splits compound terms. The concept "female leaders" in vocabulary 1 can, for example, be equivalent to "women" + "leaders" in vocabulary 2. However, the ISO standard discourages this form of mapping because it frequently results in noise. If Humord has split a compound term – whereas these are expressed as a compound term/at one class number in Dewey – the solution of several independent mappings would not work. And since it cannot be expressed with Boolean operators in SKOS, we have found no way to establish mappings between split compounds in Humord and compounds in Dewey.

ISO 25964-2 and SKOS

One of the main goals in all of our metadata development projects has been to publish both the vocabularies and the mappings in a way that is concordant with the Semantic Web. For that reason, we have from the very start had a desire to use the SKOS/RDF-standards in our work.

The SKOS-model operates with the following relationship types: *closeMatch*, *exactMatch*, *narrowMatch*, *broadMatch*, and *relatedMatch*. Additionally, the model allows for notes that are useful for definitions, explanations for how terms are used, history, comments and more. Despite the fact that SKOS is much more simplified than the ISO 25964-1-model, it in many ways successfully accommodates thesauri and includes all the relationship types we need to express. It is, of course, no coincidence that the ISO-standard and SKOS coincide in many ways; the two development milieus have followed one another closely. In the ISO-standard (p.43), this is expressed in the following way:

> As yet there is no standard schema that fully complies with this part of ISO 25964, and the development of such a schema is not within scope. However, this is a rapidly evolving field and implementers of this part of ISO 25964 should be alert to developments among interested parties, e.g. the SKOS user community.

> The schemas developed for storage purposes may also be used or adapted to enable publication of the mappings. A SKOS-compliant format […] is recommended if use in the Semantic Web is desired.

This supports our choice of SKOS and RDF.

Overarching decisions

The ISO-standard (p. 31) gives directions for which decisions need to be made at the start of every mapping project. Below follows a list of the questions that are asked and our responses to these as the pilot project was reported primo March this year. Further test mapping during this spring has given rise to a renewed discussion in the project group concerning these matters. The points of view expressed below may be altered in the transition from test mapping to actual mapping in the months to come. The text below should not be perceived as final decisions, and they should definitely not prevent an open discussion about these issues at the EDUG seminar.

Which overall model or combination of models to use:

As explained in the section "Choice of structural model", we have chosen a model[2] in which Humord is the source vocabulary and DDC is the target vocabulary. DDC serves as the *hub* in this model. Initially, Humord will be modelled with SKOS and the concepts will be mapped to Dewey classes.

How much to differentiate the mappings, in the following respects:

- whether to distinguish between equivalence and other types of mapping such as hierarchical and associative

We would like to use all three types of mapping relationships mentioned above.

- whether to accept compound equivalence mappings:

As argued above, we do not intend to use compound equivalence mappings.

- whether to distinguish between exact and inexact equivalence;

We would like to distinguish between =EQ and ~EQ.

- whether to enable establishment of more than one mapping per concept;

Since we will be mapping from a thesaurus towards a classification system with disciplinary subdivisions, we assume it will be necessary to map one and the same Humord term to several Dewey classes in the form of several independent mappings.

Whether and how to enable human mediation in the conversion process

During a previous project supported by the National Library of Norway, the Science Library at the UiO Library developed a prototype for a mapping tool, *µmapper*, in order to make the task of mapping between Realfagstermer and DDC easier and more efficient. We would like to build upon the experiences gained during this project so that it will be easier for the person performing the mapping to attain correct decisions regarding mapping candidates and relationship types. The tool developed in the present project is called *ccmapper*.

[2] This is referred to in the ISO-standard (p. 19) as "Model 3".

Intellectual challenges associated with mapping

General note: This chapter is coloured by the fact that the mapping project at the UiO Library is indeed a work in progress. As stated in the introduction, the present document is largely a translation of the pilot project report as per March 1st 2015. However, the issues in the following discussion are under constant debate and review. Some of the reasoning and points of view might be altered before we give our presentation at the EDUG seminar in April.

Mapping as a knowledge organizational phenomenon

In what respect may we say that the Humord subject heading *arkitektur* 'architecture' and the class number 720 represent the same concept? And if these two representations are to be mapped together – how should the relationship between them be expressed? These are examples of issues we have to handle when dealing with mapping. Aiming at establishing connections between concepts from two different vocabularies is an ambitious undertaking. As in all knowledge organizational practices, one has to establish a sharper distinction between conceptual categories than what we find in natural language.

A typical feature of natural language is that it is permeated by ambiguity (called polysemy) – i.e. occurrences of one term having multiple related meanings (e.g., when Norwegian *horn* signifies both 'curved parts that grow from the top of the head of some animals', a pastry ('croissant'), and a musical instrument). We also find many instances of near-synonyms (quasi-synonyms), denoting two terms which are not exact synonyms, but close enough to be considered as such in a thesaurus context (such as Norwegian *gjennomskinnelighet* 'translucence' and *gjennomsiktighet* 'transparency'). In thesauri, subtle nuances in linguistic meaning are treated by establishing distinctions between preferred and non-preferred terms (e.g., if *Norden* 'the Nordic countries' is established as a preferred term with a reference from *Scandinavia*). The most prominent example of the establishment of a sharp distinction between conceptual categories in the Dewey classification is found in the subdivision of the universe of knowledge according to a decimal structure.

The examples above indicate that knowledge organization practice offers practical solutions for handling the complexities of natural language. When facing the task of mapping, further challenges arise. The ambition is not only to connect vocabularies with different term inventories – one also has to consider the fact that different criteria for the establishment of the hierarchical structures are used. Thus, the distinctions between conceptual content may have been handled differently in each of the vocabularies. E.g., *gjennomskinnelighet* 'translucence' and *gjennomsiktighet* 'transparency' might be established as two preferred terms in one vocabulary (with a see also-reference between the terms) – and thus be regarded as two distinct concepts –while the other vocabulary might handle the same phenomenon as near-synonyms (in which one concepts is represented by two terms, of which one is preferred and one is non-preferred).

Although the purpose of mapping is to connect concepts, we are dependent on their representations in the form of terms. The hierarchical context in which a term is included, contributes to clarify the meaning ascribed to a term in a given vocabulary. Since there might have been used different criteria in the creation of these structures, one cannot simply map terms in different vocabularies on the basis of term similarity. Moreover, terms which are associated with each other in one vocabulary, might be scattered around in the structure of another vocabulary. The ambition of mapping between different knowledge structures may seem virtually impossible, unless one adopts a pragmatic attitude. This is motivated by the

desire to provide end-users with an improved subject access point to the resources in the catalogue.

In mapping subject headings with a flat structure to another vocabulary (e.g., what was done in mapping Realfagstermer to Dewey), one avoids the frustrations encountered in connection with two colliding hierarchical structures. On the other hand, there is no hierarchical context to lean on in the interpretation of terms in the source vocabulary. In establishing a methodology for mapping in this project, the point of departure is the use of Humord as the source vocabulary. However, our aim is to make the method applicable also for the mapping of other vocabularies to WebDewey.

Mapping between different structures: Thesaurus versus classification scheme

A conspicuous challenge related to mapping between a thesaurus and a classification scheme, is the fact that we are facing two different structuring principles. A thesaurus is structured in *subject* hierarchies, whereas a classification scheme organizes classes by *disciplines*. Thus, a topical domain which is found in one place in a thesaurus will be scattered on the treatment of this topic within different disciplines in a classification scheme. In using a thesaurus, one handles an inventory of terms in a hierarchical structure. In a classification scheme like Dewey, one also finds a large inventory of instructional notes, tables, and guidelines for use (i.e. the manual). This is challenging, both in relation to the selection criteria for mapping candidates in the ccmapper software tool, and for the indexer who is intended to contribute the intellectual part in the actual mapping procedure.

Based on these observations, it is a natural expectation that we will get a large amount of mappings in which one Humord concept will get two or more independent mappings to WebDewey, because the concept will be scattered throughout various disciplines in the classification scheme. It will also be relevant to map sequences of class numbers, e.g., when the general and the specific handling of a topic appear in sequential sibling numbers. An example of this is found when one is to map the Humord concept *samlinger (bibliotek)* ‘collections (libraries)’ to WebDewey, in which this topic will appear in class numbers 026 and 027, without a superordinate class number.

For compound Humord subject headings, we intend to map them to built numbers if possible, e.g., when the term *Dakota folkediktning* ‘Folk literature from Dakota’ is mapped to 398.2089975243. Usually, it is possible to class compound subject headings in either one class or by building numbers.

One possible approach in relation to mapping of subject headings which are represented by several class numbers in Dewey would be to map to class numbers for comprehensive works and interdisciplinary works. In Dewey, many topics have one class number for general treatment, as well as specific applications of the topic which is ascribed to adjacent numbers. We find an example of this when *dyr* ‘animals’ is classified in 571.1 (biology) with subordinate classes, whereas comprehensive works about animals according to instructional notes should be classified at 590. One might then consider just mapping to the class number for comprehensive works.

In the case of topics that can occur in several main classes due to various disciplinary contexts, the Dewey scheme will often provide instructional notes about where to classify interdisciplinary works. A topic like *døden* ‘death’ will occur in the scheme related to philosophy, religion, medicine, and folklore, but it will also have a class number for interdisciplinary works, within sociology at 306.9. The solution of mapping to the class

number for comprehensive works and interdisciplinary works stands out as a suitable pragmatic solution, but is still not unproblematic. The context of the subject heading in Humord may imply a different class number – e.g., the Humord subject heading *døden* 'death' is found in the hierarchy of *helse* 'health' > *kroppen* 'body' > *døden* 'death', not within sociology. On the other hand, the application of this Humord subject heading in indexing is not restricted to this aspect. Note: This perspective has been altered after the pilot project report was issued. At the present stage, we intend to map each Humord subject heading to all its locations in the Dewey schedule – class numbers and tables. The Humord context will be used to assess the conceptual content of a subject heading. However, its location in relation to academic disciplines (as expressed in the Humord structure), will not infer the principle of mapping each subject heading within all the disciplines at which this topic is represented in Dewey.

Another challenge in mapping from a thesaurus to a classification scheme, concerns the distinction between *post-coordinated* and *pre-coordinated* systems: The Humord thesaurus is intended for subject indexing for post-coordinated retrieval. The Dewey classification, in turn, is intended for pre-coordinated indexing. This causes major challenges. The stand-alone Humord subject headings intended for post-coordination will each get one or several mappings to relevant Dewey classes. When indexing using Humord, each document is assigned on average four stand-alone subject headings which will be mapped to various Dewey classes. A document on sociobiological ethics, will be assigned the Humord subject headings *sosiobiologi* 'sociobiology' and *etikk* 'ethics', which would be mapped to different main classes. In a pre-coordinated system indexed with subject heading strings, the heading 'sociobiological ethics' might have been used, which could have been mapped to a single class number in Dewey (171.7 which has a relative index term *sosiobiologi : etiske systemer* 'sociobiology : ethical systems'). It seems that it would have been easier to map pre-coordinated subject heading strings to Dewey (rather than stand-alone subject headings from a thesaurus like Humord). However, that would only be true for the cases in which the pre-coordinated elements from each vocabulary coincided.

The subject heading *morfologi* 'morphology' in Humord may illustrate how we have to consider the context of a subject heading to be able to understand which aspect of conceptual meaning we are dealing with. In Humord we find *morfologi* 'morphology' as a subordinate subject heading of *grammatikk* 'grammar'. This helps us in assessing that the relevant class number in WebDewey will be 415.9 under linguistics. Furthermore, within the topical domain of Humord it will not be relevant to map to 'morphology' associated with human or animal anatomy, which we find in two other main classes. Thus, we can establish an equivalence relationship (an exactMatch) between the subject heading *morfologi* 'morphology' and class number 415.9 – or should we? What about the consideration that by using auxiliary number -59 from table 4, you can express 'morphology' associated with the grammar of every language, and thus we can find morphology at countless class numbers? If we establish mappings to auxiliary numbers in other contexts (cf. the Dakota example), why should we not do it here? Should such mapping be done only when the subject headings in Humord are so specific that we have to use a built number to be able to express the topic in WebDewey – but not in the case where the general treatment of a topic can be expressed by a class from the main schedule (as for 'morphology')? Decision criteria for such situations are still under consideration.

The exclusion of mappings to the anatomy meanings of 'morphology' is not unproblematic in terms of the main goal of making a general Norwegian thesaurus. In such a thesaurus, every application of the concept represented by the term should be present. If we in turn map several

different vocabularies to WebDewey, we would have to maintain the original Humord context in order to justify an equivalence relationship from *morfologi* 'morphology' to 415.9, since this type of relationship is presupposed to be reciprocal between the vocabularies.

Structural similarities between Humord and a classification scheme organized by disciplines

In addition to the different organizing principles that underlie thesauri and classification systems, we must expect that any knowledge organization system will be coloured by origin, i.e. decisions and solutions along the way. In our case, we see for example that Humord contains deviations from the thesaurus construction principles, and also that Dewey is characterized by choices made during its long history. Examples in connection with Humord are that we have a finite set of top terms which resembles a division of disciplines – which is understandable considering that Humord originates from subject indexing initiatives in several department libraries. Furthermore, there might have been used different criteria for sub-division within each hierarchy. In Dewey, we find elements of faceted structure (e.g., biology and music) in a predominantly enumerative schedule. This makes our efforts in establishing criteria for selection of mapping candidates from Humord and WebDewey even more challenging. We cannot assume that a pattern found in connection with test mapping of one topical domain in Humord, will appear in the same way elsewhere in the thesaurus.

The question as to what extent we should take into account how a thesaurus term has been used according to the bibliographic data, may also be illustrated by the subject heading *hester* 'horses', which in Humord is located in the following hierarchy: *realfag* 'science' > *naturvitenskap* 'natural science' > *biologi* 'biology' > *zoologi* 'zoology' > *dyreliv* 'wildlife' > *dyr* 'animals' > *vertebrater* 'vertebrates' > *pattedyr* 'mammals' > *hester* 'horses'. Yet this subject heading is used in bibliographic records for documents concerning horses as a motive in art. In a general thesaurus, there should be a rich amount of top terms to avoid such dilemmas – and the subject headings should not be organized under hierarchies of disciplines. In Humord we find the reverse situation, when all of the 18.500 subject headings are submerged under 26 top terms. This must necessarily cause major challenges. Structurally, Humord is very similar to a classification scheme, but it is applied as a thesaurus – with stand-alone subject headings intended for post-coordinated indexing.

The dilemmas that arise in connection with the academic orientation of Humord (subject headings organized according to disciplines) – contrary to the principle of subject orientation in thesaurus construction – can be further elucidated using the example *tobakk* 'tobacco'. This topic is a favourite example for demonstrating how a topic can be scattered throughout different disciplines in a classification scheme. In Dewey, the topic 'tobacco' is spread throughout various disciplines all over the schedule – at class numbers for botany, ethics , religion , agriculture, human toxicology, production technology, customs, smuggling, etc. In Humord we find the subject heading *tobakk* 'tobacco' in the following hierarchy: *næringsliv og økonomi* 'business and economics' > *produksjon* 'production' > *produkter* 'products' > *nytelsesmidler* 'stimulants' > *tobakk* 'tobacco'. Does it follow from this that we should only map this concept to the WebDewey class number for 'tobacco' in relation to production technology?

We often have to deal with discrepancies between the Humord structure and how the vocabulary is used in indexing. The 'tobacco' example shows that in mapping, we have to choose whether to take into account both of these aspects, or either one of them. Should we primarily take into account that Humord has an overall academic structure (which is contrary to the principles of thesaurus construction) – or do we focus on the fact that the vocabulary is

used in indexing like a thesaurus (without considerations of how the subject headings are related to academic disciplines in the hierarchies)? What effect will the decisions relative to these questions have in terms of end-user perspective?

If we have a look at the bibliographic records for documents which have been indexed with the subject heading *tobakk* 'tobacco' in Humord, we realize that they are classified at Dewey numbers from all over the schedule. Should we then map to all these classes? In the morphology example above, we used the topical domain covered by Humord as a criterion for selecting one mapping rather than potentially three. In the case of 'tobacco', we cannot say that the topical domain for Humord is restricted to the production aspect of tobacco (even though the subject heading is found under production of stimulants). Should we then map to all the class numbers except for science and medicine (which would be covered by Realfagstermer and MeSH) – i.e., a selection of the aspects listed above? In discussing these types of issues, we acknowledge that the structure of Humord – with a very small number of top terms – "forces" terms into hierarchies which are not reflected in actual indexing practice, according to the bibliographic data. We also acknowledge that we cannot come far enough with statistical mapping. It is absolutely necessary to consider the context both in the source and the target vocabulary.

Mapping candidates: Selection of pairs of concepts to be mapped, and consideration of relationship types

As mentioned earlier, the ISO standard 25964-2 presupposes three logical types of mappings (equivalence, hierarchical, and associative), labelled with five relationship types: exact equivalence, inexact equivalence, broader mapping, narrower mapping, and related mapping. In the following sections, the SKOS labels for these relationships will be used: skos:exactMatch, skos:closeMatch, skos:broadMatch, skos:narrowMatch, and skos:relatedMatch. Let us look at some examples of how mappings between Humord and WebDewey can manifest themselves – first, in relation to the assessment of which pairs of elements from the two vocabularies should be mapped.

We saw earlier that the Humord subject heading *samlinger (bibliotek)* 'collections (libraries)' might be mapped to the sequential numbers 026-027 in Dewey, i.e. the class numbers for collections in general libraries (027) and collections on specific subjects (026), respectively. So this is an example of a topic which appears in several class numbers in Dewey. We then get a case in which one Humord subject heading gets two independent mappings to WebDewey. You can then either establish a mapping to the number sequence 026-027, or you can create an independent narrowMatch to each of these class numbers. The Humord subject heading is more general than each of these classes, and moreover, each of the numbers we map to has further subdivisions in WebDewey. An alternative solution to multiple independent mappings of the subject heading *samlinger (bibliotek)* 'collections (libraries)', would be to establish a broadMatch to class 020, tagged with the caption *bibliotek- og informasjonsvitenskap* 'library and information science'.

If we consider the Humord subject heading *normativ etikk* 'normative ethics', it is easy to find a link to WebDewey class 170.44, captioned by 'normative ethics'. The hierarchical contexts of the subject heading and the class number correspond well, so we can establish an equivalence mapping – but should it be exactMatch or a closeMatch? If we are to use exactMatch at all in mapping, this will be a good example of such a case. You just have to be aware that aspects of the conceptual content assigned a class number also can be embedded in other class numbers, for instance: Class number 170.44 applies to comprehensive works on normative ethics, while normative ethics in specific contexts would be classified with that

specific aspect, e.g. *yrkesetikk* ‘occupational ethics’ at 174. This is a recurring theme in association with mapping between a thesaurus and a classification scheme.

In the same hierarchy as *normativ etikk* ‘normative ethics’ in Humord, we find *sinnelagsetikk* ‘ethics of conviction’, which we do not find in WebDewey. In this case we will probably have to establish a mapping to 170.4, i.e. *spesielle emner innen etikk* ‘special topics of ethics’. In isolation, this would vouch for a broadMatch, because the concept in the target vocabulary is more general than in the source vocabulary. At the same time, one has to overlook that *sinnelagsetikk* ‘ethics of conviction’ (due to the lack of subdivision at this WebDewey class) thus will get a class number in Dewey which is superordinate to *normativ etikk* ‘normative ethics’. This happens even though the concepts in question should rather have appeared on the same hierarchical level, and should have been represented as sibling class numbers. Anyway, this example demonstrates that when Humord subject headings are mapped to WebDewey and are made available from the same interface, those parts of Humord which have a higher granularity than the Dewey table, will enrich WebDewey with specific terms that will be useful, both for end-users and in document indexing.

The subject heading *heraldikk* ‘heraldry’ in Humord may illustrate a recurring challenge encountered in our project, i.e., the choice between closeMatch versus broadMatch. The term *heraldikk* ‘heraldry’ has the following entry in Humord:

> *heraldikk* ‘heraldry’
> *BF slektsvåpen* ‘UF family weapons’
> *BF våpenskjold* ‘UF coats of arms’
> *OT kulturkunnskap* ‘BT cultural studies’
> *TT kulturkunnskap* ‘TT cultural studies’
> *UT emblemer* ‘NT emblems’
> *UT faner* ‘NT banners’
> *UT flagg* ‘NT flags’
> *UT riksvåpen* ‘NT national coats of arms’
> *SO ordener (utmerkelser)* ‘RT orders (honorary)’

In WebDewey we find *heraldikk* ‘heraldry’ in class 929.6:
900 Historie og geografi ‘history & geography’
920 Biografier og genealogi ‘Biography & genealogy’
929 Genealogi, navn, insignier ‘Genealogy, names, insignia’
929.6 Heraldikk ‘Heraldry'
Inkluderer: Våpenmerker; byvåpen, kommunevåpen ‘Including crests’
Her: Våpenskjold ‘Class here armorial bearings, coats of arms’

In the mapping process we have to ask ourselves whether the Humord term *heraldikk* ‘heraldry’ represents the same concept as class 929.6 in WebDewey. At the outset, this may appear to be a simple case of exactMatch, since the Humord subject heading coincides with the full caption in WebDewey. However, there turns out to be several complicating factors:

If we look at the superordinate hierarchical context, *heraldikk* ‘heraldry’ in Humord is a narrower term to *kulturkunnskap* ‘cultural studies’, whereas heraldry in class 929.6 in WebDewey is found under the discipline *genealogi* ‘genealogy’, which again is found under the discipline *historie* ‘history’. In Humord, *genealogi* ‘genealogy’ is found in another hierarchy than *heraldikk* ‘heraldry’, viz. as an alternative term to *slektsforskning* ‘genealogical research’, which we find under *historieforskning* ‘historical research. If we look at the

subordinate hierarchical context to *heraldikk* ‘heraldry’ in Humord, we find, among other things, *faner* ‘banners’ and *flagg* ‘flags’. In WebDewey these topics are found at 929.92, i.e. as a subordinate number of a sister class to 929.6, where we found *heraldikk* ‘heraldry’. May we consider the Humord subject heading *heraldikk* ‘heraldry’ and class number 929.6 as different representatives of one and the same concept? To what extent do the represented concepts coincide, and which mapping relationship type would be relevant to use in this case?

As we saw in the morphology example above, for mapping to be possible, we have to use contextual information of thesaurus terms to assess their conceptual content. Based on the Humord context for *heraldikk* ‘heraldry’, we can deduce that we are talking about crests and symbols of kinship. Thus, we may disregard the other application of *heraldikk* ‘heraldry’ found in WebDewey, as used in connection with the class number representing *kongehus, adel og ridderordener* ‘royal houses, peerage, and orders of knighthood’ at 929.7 – which has the relative index term *arverekkefølge (heraldikk)* ‘precedence (heraldry)’. We must also choose to ignore the fact that the subordinate terms in Humord can be found at different classes in WebDewey. E.g., *faner* ‘banners’ and *flagg* ‘flags’ are not only found at 929.92, but also with military science at 355.15.

Using the context of a Humord subject heading to understand the conceptual content of a term does not, however, mean that we can assume that the superordinate and subordinate hierarchy of this term and the corresponding class number should match – this is a given when mapping between a thesaurus and a classification scheme. So here we will choose to make a mapping between the subject heading *heraldikk* ‘heraldry’ and class number 929.6 – but which relationship type should be established? The contexts of the source and target vocabularies preclude an exactMatch, so we face a choice between a closeMatch and a broadMatch. Two of the subordinate terms to the Humord subject heading are not found in WebDewey (*emblemer* ‘emblems’ and *riksvåpen* ‘national coats of arms’), whereas the other two are found on another class number (*faner* ‘banners’ and *flagg* ‘flags’ at 929.92). None of the terms in the including note at 929.6 occurs as Humord subject headings. Still, this is obviously the relevant class number to use for this mapping. Since we cannot ascertain a generic relationship between the concepts in the two vocabularies, we choose to establish a closeMatch equivalence relationship.

Test mapping

In order to reveal the complexity which will emerge in connection with the actual mapping, we are performing a number of test mappings in our project. The examples in the preceding sections are garnered from this work. The experiences we accumulate during the test mapping will be integrated as algorithms in the mapping tool. Issues which cannot be solved using automated methods must be established as guidelines for consistent practices for the intellectual contribution in mapping work. In the test mapping procedure, our first step is to search for relevant mapping candidates for a Humord subject heading. We then uncover how different assessment criteria for the selection of candidates will affect what mapping suggestions you get (e.g., making several independent mappings of one subject heading, versus mapping to a superordinate class number). We also consider which relationship type should be established for each mapping. Several dilemmas arise in this process.

Let us illustrate with the following example: The Humord subject heading *samarbeid* ‘cooperation’ occurs in the following hierarchy: *samfunnsvitenskap* ‘social science’ > *sosiologi* ‘sociology’ > *sosiale prosesser* ‘social processes’ > *samarbeid* ‘cooperation’, with the subordinate terms *gruppearbeid* ‘group work’, *internasjonalt samarbeid* ‘international cooperation’, and *regionalt samarbeid* ‘regional cooperation’. When we map *samarbeid*

'cooperation' to Dewey class 302.14 – which has the caption *sosial deltakelse* 'social participation' – should we then establish a broadMatch or a narrowMatch? If we consider the superordinate context for the Humord subject heading (i.e., 'social processes', 'sociology', and 'social science'), then *samarbeid* 'cooperation' can be interpreted as narrower than the caption *sosial deltakelse* 'social participation', and we will get a broadMatch from the Humord subject heading to the class number. If, however, we consider the subordinate context of the Humord subject heading (i.e. *gruppearbeid* 'group work', *internasjonalt samarbeid* 'international cooperation', and *regionalt samarbeid* 'regional cooperation'), then *samarbeid* 'cooperation' can be interpreted as broader than *sosial deltakelse* 'social participation', and we will get a narrowMatch.

We have demonstrated that the test mapping procedure concerns the choice of which class numbers each Humord subject heading should be mapped to, as well as considerations regarding relationship types. To be able to feed these conditions as algorithms into the mapping tool, an assessment of the degree of conceptual similarity between each Humord subject heading (including context) and each mapping candidate class number (with its context in WebDewey). We also need to evaluate in which situations we consider other sources (as, e.g., UiO Library Subject Index to Dewey, as well as bibliographic data) to provide a useful context – versus noise – in the mapping tool.

The intellectual contribution in computer-assisted mapping

As already stated, we are aiming at developing a methodology for mapping as a computer-assisted intellectual task. This ambition requires a tool which can be used for decision-making in two kinds of choices: Firstly, which subject headings in Humord should be mapped to which class numbers in WebDewey? Secondly, which relationship type should be established for each mapping?

Let us examine the first question, concerning which items should be paired: Ideally, the mapping tool should come up with relevant mapping candidates, presented according to decreasing relevance. When this is not always possible, the candidate list will contain noise: The indexer, who performs the intellectual contribution in the mapping task, will sometimes have to consider many irrelevant mapping suggestions. At the same time, the absence of relevant pairs of mapping candidates poses a great challenge: If the indexers are to be able to notice that relevant mapping pairs are missing in the list of candidates, this will require a thorough knowledge of the Dewey classification scheme.

A main issue in our project is to explore the potentiality of automated methods in mapping. Where do we find the boundary between what is serviceable to automate, and which operations will anyway require an intellectual assessment? An example of a task which is difficult to automate, is mapping to built numbers. These will need to be built manually. On the other hand, this would provide a very useful contribution, considering that one of the intentions of the project is to present the results of the mappings as Humord subject headings in WebDewey. The Norwegian WebDewey will thus be enriched with a large number of pre-built numbers within the topical domain covered by Humord.

Regarding the choice of relationship types, this will have to be an intellectual operation. The contribution from the mapping tool will be to present the context for the mapping candidates in the source and target vocabularies. With this in mind, we attempt to clarify which elements from each of the vocabularies that provide us with the best basis for our decisions. Too much input will only lead to mental overload for the indexers.

According to the ISO standard, the equivalence relationship (i.e. exactMatch and closeMatch, in SKOS terms) will be the most frequent relationship assigned in mapping. This is probably more applicable when mapping between two thesauri, rather than between a thesaurus and a classification scheme. So far in our project, exactMatch seems to be rarely applicable, with the exception of examples like the top term *samfunnsvitenskap* 'social science' in Humord, which can be mapped to main class 300 in Dewey. The class numbers often represent a cluster of concepts, and accordingly an exactMatch will be irrelevant. However, we are often faced with the problem of whether to establish a closeMatch or broadMatch relationship. The decision concerning relationship types will have consequences for subsequent utilizations in retrieval, e.g., if one would like to exploit established mappings in automated query expansion. These are issues which we would like to discuss at the next seminar in the EDUG collaboration. The mapping seminar in April is aimed at establishing joint guidelines for best practices in mapping, from an end-user perspective.

In the preceding discussion, we have seen examples of equivalence (exactMatch and closeMatch) and hierarchical relationships (broadMatch and narrowMatch) – but when will an associative mapping relationship (relatedMatch) be relevant? According to the standard, one may establish an associative relationship if the concept in the source vocabulary can be associated with the concept in the target vocabulary in such a manner that on may assume that the target concept (in our case represented by a class number) will be relevant for a user searching for the source concept (the Humord subject heading). The ISO standard demonstrates with the example of the relationship between 'e-learning' and 'distance education'. At the same time, the standard suggests that there will be a blurred distinction between a relatedMatch and a closeMatch, and states that pragmatic choices have to be made in relation to the end-user perspective.

So far in the test mapping procedure, we have used relatedMatch only when establishing a mapping to the class number for related terms (i.e. see also-references to a preferred term in Humord). In the aforementioned example of *samlinger (bibliotek)* 'collections (libraries)', for which a mapping was established to class numbers 026 and 027, it could be relevant to consider a relatedMatch to 025.21 *samlingsutvikling* 'collection development'. This concept belongs to a different conceptual category (i.e. operation versus object), but this mapping might be useful for end-users searching for *samlinger* 'collections'.

We have seen that the ISO standard presupposes a distinction between five mapping relationships, but that the actual task of mapping raises several issues when it comes to the choice of relationship type in each case. Progress in these matters requires practical guidelines for mapping. Without consistent mapping practices, we will lose the potential benefit of applying different mapping relationship types in end-user tools. Further test mapping will make clear whether it might be useful to apply a lower amount of relationship types, e.g., only closeMatch and narrowMatch – or even just establish non-differentiated mappings.

Both elements of the intellectual contribution in computer-assisted mapping proves to be very challenging – i.e. both choosing which elements are to be connected, and how they should be related. The purpose of the mapping tool is to provide a relief (rather than a load) in the selection process of mapping candidates and relationship types. When we have to make pragmatic choices to make mapping possible, the aim is to be guided by an end-user perspective: Which outcome will provide for the most effective subject access to the information resources?

Design of the mapping tool ccmapper

General note: In this chapter we present *ccmapper*, i.e. the mapping tool that is under preparation at UiO. Depth reading of the technical details is unnecessary for readers who only need to get a general picture of the information architecture for the software.

As mentioned in the chapter on intellectual challenges associated with mapping, we envision mapping as a computer-assisted intellectual task. As such, the mapping tool will support the endeavour in two main ways:

1. By suggesting the most relevant mapping candidates for a given subject heading
2. By providing a good overview of the contexts associated with concepts in both the source and target vocabularies

The current project is based on experiences garnered in a previous project, in which a source vocabulary, Realfagstermer, was mapped to the preliminary Norwegian translation of the Dewey Decimal Classification (DDC). During this project, a prototype tool named µmapper was developed (see illustration, figure 2):

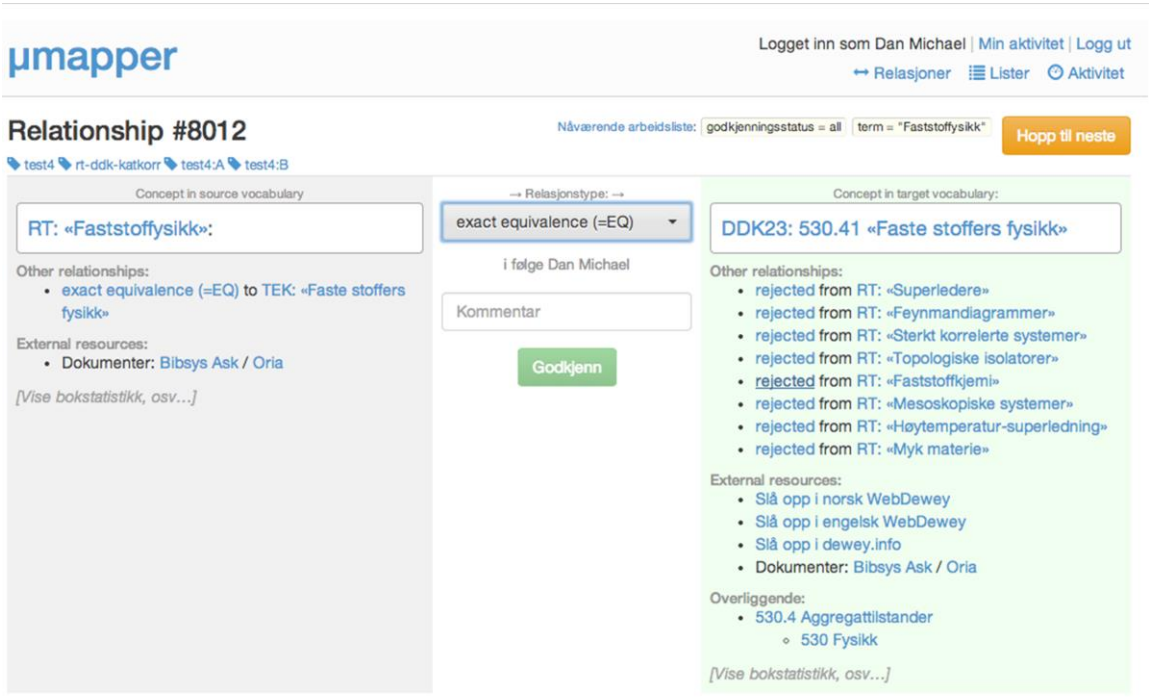

Figure 2: Screen-shot from µmapper

As mentioned earlier, Realfagstermer is a list of subject headings with a flat structure. Humord, the source vocabulary in this project, is on the contrary a thesaurus with a hierarchical structure.

Automatic mapping suggestions in µmapper were based on term equivalence between headings in Realfagstermer and DDC. Additionally, statistical mapping was used in an attempt to find correlations between Realfagstermer and DDC classes in bibliographic records, but the data source was too sparse to generate meaningful suggestions. Literary warrant played, however, an important role in the intellectual assessment of the meaning behind a subject heading or DDC class, where this was necessary. The assessment of the

target concept's placement in the Dewey hierarchy was also intellectual, but the target vocabulary was limited to the 500 and 600-640-groups in DDC, which made the assessments more manageable than if all DDC classes had been used in the target vocabulary.

Compared with Realfagstermer, Humord has more contextual information that can be taken advantage of to generate good suggestions. For example, each Humord has at least one index term and one hierarchical relationship, and may include alternative terms and notes. This creates new challenges to take into consideration in the current project.

A Dewey class, on the other hand, represents a concept that is represented by a notation, and that comes into expression through its class heading, its hierarchical context, alternative headings (including synonyms) and class notes.

> Mappings to or from a class or category in a monohierarchical scheme should treat the class/category as a pre-coordinated concept whose meaning can be established by inspecting all its superordinate and subordinate classes as well as any scope notes associated with it. Inspection of the caption alone is inadequate.
>
> (International Organization for Standardization, 2013, s.32)

Early in the current project, it became clear to us that we needed a more advanced mechanism than traditional term comparison as the basis for suggestions in our mapping tool. To come up with good mapping suggestions, it was deemed advantageous that the mapping tool deals with concepts instead of terms, primarily by taking the contexts of both the subject heading and DDC class into consideration.

ISO 25964-2, chapter 14, "Techniques for identifying candidate mappings" and chapter 14.2 "Computer assisted direct matching" describe the suggested approach and accompanying requisite to have an as good as possible overview of both the source and the target vocabularies:

> The candidate mappings identified by the matching processes described should be assembled for review by an expert. For each concept in the source vocabulary, the expert should be able to view the complete record (including scope note, broader and narrower terms).
>
> […]
>
> The viewing interface should make it easy to check the complete context of each concept identified in the target vocabulary. It should also support the expert in selecting the appropriate type of mapping for the candidate(s) he approves.
>
> (s.40)

The mapping tool should present suggestions in the form of a list of mapping candidates, which in turn will be intellectually assessed by an expert. The expert decides which mapping relation to use for the desired mapping candidate and rejects the other candidates.

For each mapping candidate, the tool must give the expert a compact and best possible overview of the source and target concepts' meanings and contexts as support mechanisms for the assessment of the mapping relation.

We have chosen to use the user interface metaphor *dashboard* in designing the tool. Figure 3 on the next page illustrates our preliminary ideas, as of February 2015. Good interaction design will be an important factor in success.

The intended user group for the tool is librarians. We can therefore allow for a tool with a more complex user interface than we would for a regular web app. The users of the tool have a background in library and information science and will receive training.

In addition to the source and target vocabularies, we intend to take advantage of other data sources that can either provide better context for the concepts to be mapped, or give statistical information about the actual use of the subject headings with Dewey classes in bibliographic records. We plan to use the UiO Library Subject Index to Dewey, as well as co-occurrence in bibliographic records.

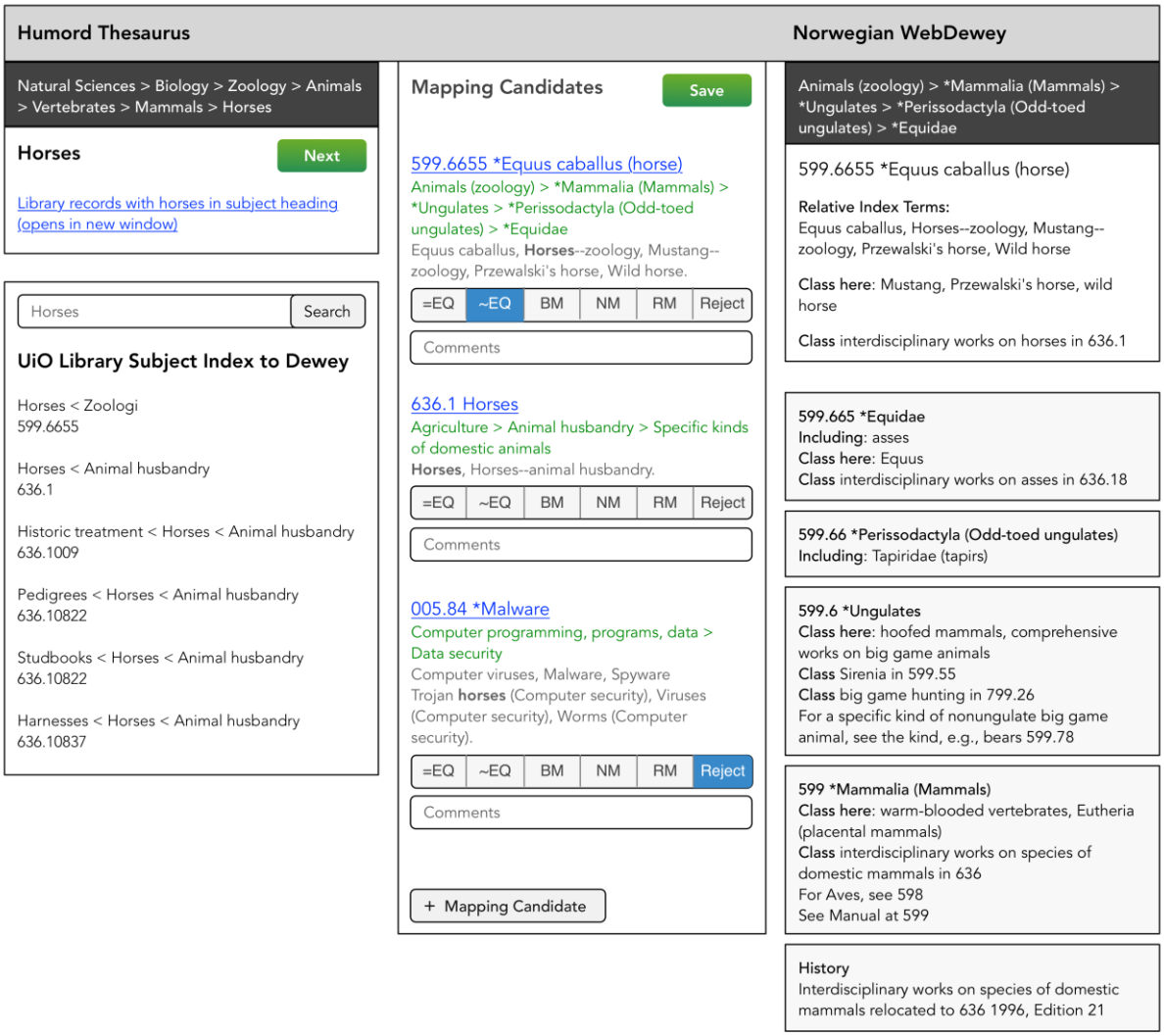

Figure 3: Interaction design for the prototype ccmapper, as of February 2015.

The tool's name, *ccmapper*, stands for *concept context mapper*. We plan to have a finished prototype of the tool in the course of the spring, 2015.

Below follows a description and discussion of various challenges associated with the realisation of the mapping tool.

Data sources, SKOS and Linked Data

As part of our mapping preparations, we have gathered and converted various data sources we intend to use in the tool:

- The thesaurus Humord
- The Norwegian WebDewey
- The UiO Library Subject Index to Dewey
- Bibliographic records from the four sections of the UiO Library using Humord

All data sources have during the course of the project been converted to SKOS and place in a GIT repository[3] for version management.

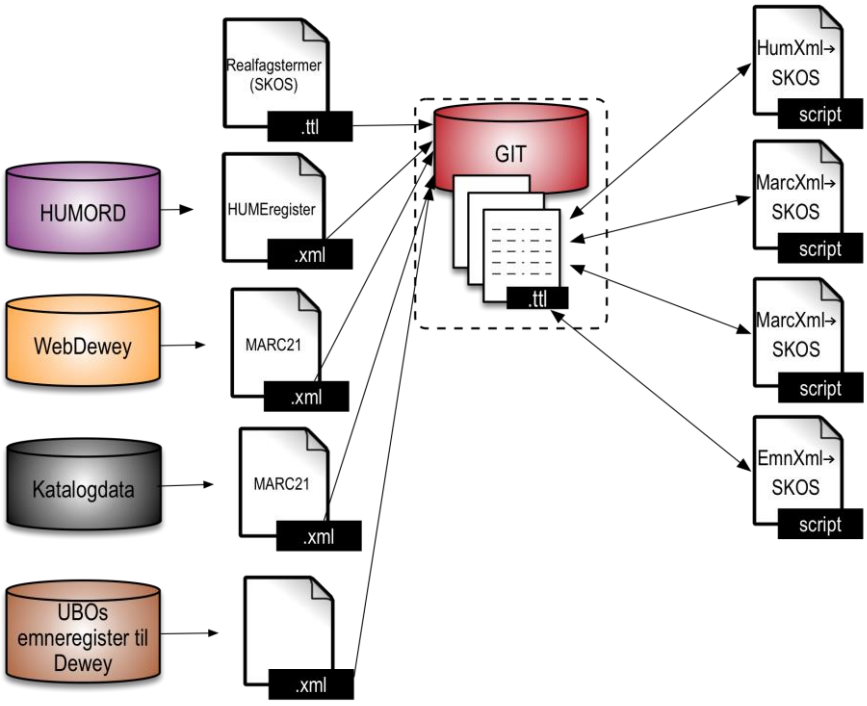

Figure 4: Data sources

Most of the data sources are published openly with a CC0 1.0-license[4]. Source files and converted files for the Norwegian WebDewey are not openly published due to OCLC license restrictions.

Humord, the UiO Library Subject Index to Dewey and Realfagstermer are published openly as Linked Data at data.ub.uio.no.

We expect to do continuous adjustments to the code based on experiences in this project. We must also consider developments in the cooperative project *BIBSYS Library Database in the Semantic Web*, which has received funding from the National Library of Norway for 2015, regarding the publication of data in bibliographic records as Linked Open Data. This project is a cooperation between BIBSYS and the University of Oslo (UiO), the University of Bergen, the Norwegian University of Science and Technology, and University of Tromsø, the Arctic

[3] https://utv.uio.no/stash/projects/UB/repos/datakilder/
[4] http://creativecommons.org/publicdomain/zero/1.0/

University of Norway. One of the goals is to develop one common RDF-representation of bibliographic data instead of several local variants.

We must also consider when and how the data sets will be published in the long run. If the National Library of Norway is the owner of a national thesaurus, the data should be published within the National Library's domains, for example at data.nb.no or data.norge.no.

SKOS (Simple Knowledge Organization System)

Since we are working with data from various knowledge organization systems, the W3C-standard SKOS works well as a common data model. SKOS is an RDF-vocabulary, and coding of the data sources as Linked Data opens up the possibility for better searching and connecting of data across various systems.

SKOS was created to enable modelling of regular, but relatively simple and semi-formal, knowledge organization systems like thesauri, taxonomies, classification systems, folksonomies and controlled vocabularies.

The standard also includes relation-types for mapping.

> SKOS occupies a position between the exploitation and analysis of unstructured information, the informal and socially-mediated organization of information on a large scale, and the formal representation of knowledge.[5]

In order to be general, SKOS has an "Emphasis on minimal ontological commitment" and is intended to be a link between more informal information structures and ontologies that may be based on logic and axioms.

The main SKOS-model to model knowledge organization systems is relatively simple (not including the mapping relation-types):

- Modelling of entities
  - Concept skos:Concept
  - Label (terms referring to concept):
    - Preferred Label skos:prefLabel
    - Alternative Label skos:altLabel
    - Hidden Label skos:hiddenLabel

- Modelling of relation-types
  - Broader/Narrower skos:broader, skos:narrower
  - Associative relation skos:related

  - Documentation, i.e. various types of notes
  - skos:scopeNote, skos:definition, skos:example, skos:historyNote, skos:editorialNote, skos:changeNote

There are five mapping relation-types in SKOS, as discussed earlier in this report and illustrated in the interaction design for the prototype in figure 3.

[5] SKOS W3C Recommendation, kap. 1.1, Background and Motivation.http://www.w3.org/TR/skos-reference/#L879

ISO 25964-2	Abbreviation	SKOS
Exact equivalence	=EQ	skos:exactMatch (symmetric, transitive)
Inexact equivalence	~EQ	skos:closeMatch (symmetric)
Broader mapping	BM	skos:broadMatch (inverse of narrowMatch)
Narrower Mapping	NM	skos:narrowMatch (inverse of broadMatch)
Related Mapping	RM	skos:relatedMatch (symmetric)

Data sources

As part of our mapping preparations, we have gathered and converted data from the various data sources we intend to use.

Realfagstermer

Realfagstermer already existed at project start as RDF/SKOS with a CC0-licence. In the course of the fall, 2014, the operational plan for the vocabulary has been moved to data.ub.uio.no so that the open data is always updated.

Humord and the UiO Library Subject Index to Dewey

Humord and the UiO Library Subject Index to Dewey were not available as open data at the start of this project. Both vocabularies are maintained in the EMNE-module in BIBSYS, which supports XML-exports. We have converted both vocabularies to RDF/SKOS and published the datasets with a CC0-licence at data.ub.uio.no. An example term from Humord in SKOS (and Turtle-serialization) looks like this:

```
<http://data.ub.uio.no/humord/c05316> a skos:Concept ;
   dct:identifier "HUME05316" ;
   dct:modified "1994-03-21"^^xsd:date ;
   skos:altLabel "Bildende kunst"@nb,
       "Billedkunst"@nb,
       "Visuell kunst"@nb ;
   skos:broader <http://data.ub.uio.no/humord/c05183> ;
   skos:definition "Bildekunst omfatter tradisjonelt visuell kunst: malerkunst,
tegnekunst, grafisk kunst, bildehoggerkunst og bildevev. Her også nye medier som
fotokunst, videokunst mm <UBB>"@nb ;
   skos:inScheme <http://data.ub.uio.no/humord/> ;
   skos:prefLabel "Bildekunst"@nb .
```

With the conversion, we have gone from a term-based model in the tradition of ISO 2788 to a concept-based model where each term has a preferred term (skos:prefLabel) and zero or more alternative terms (skos:altLabel). See-references are converted to alternative terms (*sykkelstier SE sykkelveier* 'bike paths SEE bicycle paths' becomes skos:prefLabel "sykkelveier"; skos:altLabel "sykkelstier").

The only real challenge in this modelling transformation comes in the form of general see-references (split non-preferred terms): references from one term to two concepts (for example, the reference *buddhistisk filosofi, SE buddhisme * filosofi* 'buddhist philosophy, SEE buddhism * philosophy' in Humord). If these are to be expressed in RDF, the result diverges from the SKOS-model (the ISO 25964-expansion of SKOS delineates one such option). General see-references are an internal usage convention in the Humord thesaurus and would either way never be mapped to Dewey.

Other challenges regarding conversions to RDF/SKOS include facet indicators and guide terms, but for the purpose of mapping, neither of these classes are interesting because they represent technical elements of the thesaurus that are not used in indexing. We have chosen to include them in the conversion for display purposes, but have classed them differently so that they can easily be distinguished from regular subject headings.

Norwegian WebDewey

The Norwegian WebDewey is not finished, but the National Library of Norway has given us access to export status quo from the translation software in the MARC21XML-format for internal use. These have been converted to RDF/SKOS with a few local expansions. Building on the model found at dewey.info, we have included relative index (as skos:altLabel) and notes. Four types of notes are especially interesting because they include references to terms that are clearly demarcated in their own subfields:

Note type	Example in MARC21 Classification	Example converted to RDF
"Andre betegnelser" (Variant-names, ess=nvn)	680$iAndre betegnelser:$tVaskulæare kryptogamer$i,$tvaskulære planter uten frø$9ess=nvn	wd:variantName "Vaskulæare kryptogamer"@nb, "Vaskulære planter uten frø"@nb
”Her” (Class-here, ess=nch)	680$iHer:$tAssosiative algebraer$i,$tikke-kommutative algebraer$t,ikke-kommutative ringer$9ess=nch	wd:classHere "Assosiative algebraer"@nb, "Ikke-kommutative algebraer"@nb, "Ikke-kommutative ringer"@nb
"Inkluderer"-noter (Including, ess=nin)	680$iInkluderer:$tKorrekturlesing$9ess=nin	wd:including "Korrekturlesing"@nb
"Tidligere klassebetegnelse" (Former heading, ess=nph)	680$iTidligere klassebetegnelse:$tKidneybønner$9ess=nph	wd:formerHeading "Kidneybønner"@nb

More descriptive notes, like definition notes (definition note, ess=ndf) and scope notes (scope notes, ess=nsc) are also converted, but we are as yet unsure if these can be exploited in mapping.

The mapping schema and description can be found at https://github.com/scriptotek/mc2skos. We do not have permission to publish the actual data.

Indexing and classification data from bibliographic records

For each pair of concepts from diverse vocabularies (for example Humord and DDC) we can find a statistical degree of association based on how often the pair occurs together compared to how often each member of the pair occurs alone in the bibliographic records. From an earlier project, we have access to a data dump of bibliographic records up to April 2014, and we intend to ask BIBSYS for a new data dump when we get closer to the beginning of our mapping work.

In the bibliographic records, there are Dewey class numbers from many different editions. Ideally, we could limit our analysis to one specific edition, like DDC-23, but we only have ~ 50 000 records that contain both DDC-23 and Humord and ~ 7 000 records that contain both

DDC-23 and Realfagstermer. Relative to the sizes of the relevant vocabularies, these numbers are too low. To increase the size of the data set, we have opted to limit by time. If we for example limit to bibliographic records starting with the year 2000, we get ~ 200 000 records with both Humord and DDC and ~ 15 000 records with Realfagstermer and DDC. For Realfagstermer, the data set is still small, but for Humord, it will be interesting to see if the statistical mapping can provide good suggestions.

Automatically generated list of mapping candidates

We have taken a closer look at several tools used to solve similar issues[6]. It appears as though there are very many related concepts, approaches and disciplines that partially overlap.

Within Linked Data, the concepts "Equivalence Mining", "Equivalence Matching" and "link discovery" are established. Here, tools like Silk ("A Link Discovery Framework for the Web of Data") and LIMES ("Link Discovery Framework for Metric Spaces") appear to be the most used.

When it comes to linking data sets that do not use common identifiers, in particular database records in diverse systems, the following concepts are used: "Record linkage", "Entity resolution", "Name resolution", "Identity resolution", "Deduplication" and "Merge/purge".

Common for the tools and approaches we have analysed is that they are largely based on the comparison of term equivalence.

Other related approaches and disciplines include amongst others "Ontology Mapping", "Ontology Alignment", "Ontology Matching", "Semantic Matching" and "Semantic Mapping". *Ontology Matching* (Euzenat & Shvaiko, 2013) offers a systematic summary over known techniques and strategies.

We would like a data-generated list with suggestions for mappings that in a flexible way takes the contexts of both the term and the class into consideration. We would also like to be able to easily change the weighting of various forms of context, and the list should be ordered by relevance.

We have therefore chosen to use vector space modelling, which allows us to take all forms of context into consideration, regardless their weighting, in an easy and standardized method.

For each concept in the source and target vocabularies, we plan to generate a synthetic text document to represent the concept. The document will include textual representations of everything that makes up the content and context of the concept, and all elements can be weighted by repeating them a certain amount of times.

By using the vector space model (Salton, Wong, & Yang, 1975) already ubiquitous in modern indexing and search, we can use standard search technology like Apache Lucene[7] to index and search in the documents.

For each subject heading, we can use the accompanying term vector as search vector and order the term vectors from Dewey documents by the cosine of the angle distance to the term vector for the subject heading (i.e. standard vector-based search).

---

6 http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining
7 http://lucene.apache.org

In addition, we will use TF-IDF (Term frequency-Inverse document frequency) to weight the terms. TF-IDF gives high weightings to words that occur rarely in the document collection. Basically, this means that words that help to distinguish between various concepts are weighted higher than words that occur frequently (for example, in many different Dewey classes) and therefore do not help to distinguish the concepts from one another.

We realize that we will need to normalize the terms in the index, that is to say, lemmatize them, in order to get better matches for the subject heading to the right Dewey class. We would therefore like to use a newly developed lemmatization module for Lucene/solr that uses Språkbanken's (the National Library's language technology resource collection for Norwegian language) word lists.

Additionally, we see that in a number of instances, it will be necessary to split compound words in order to get the correct match. The article *Monolingual document retrieval for European languages* (Hollink, Kamps, Monz, & De Rijke, 2004) shows that compound splitting for Swedish increases mean average precision by 25 % compared to standard lemmatization.

We are, however, unsure as to whether compound splitting can be achieved in this project.

A short summary of the methodology:

- Use term and available context to create a synthetic text document which serves as a representation of the concept to be mapped
    - Filter out stop words
    - The various components of the synthetic document must be weighted. (The preferred term, for example, must be weighted higher than hierarchically broader terms)
    - We can include context from related systems in the synthetic document, for example from the UiO Library Subject Index to Dewey and from co-occurrences of subject heading and Dewey classes in bibliographic records.
- Index the text documents with Lucene and use tf–idf (term frequency–inverse document frequency)[8]
    - Eventual splitting of compound words (add splitted terms to index)
    - Stemming/lemmatization of words to get uniform word stems.
- Use document vectors for the specific subject heading as the search vector in searches against the document index for Dewey classes
- Order the results list by cosine-similarity for the angle between the vectors

Use of the subject index and record data as context

Data from the UiO Library Subject Index to Dewey will be used to give context within the user interface, as explained in the interaction design figure above.

The UiO Library Subject Index to Dewey is developed by use of the chain indexing method and therefore puts terms into their disciplinary context. When we find co-occurrences between a term in Humord and the subject index, the link between the subject term and the Dewey class number will offer a strong candidate for mapping. We will therefore include class numbers in the synthetic document.

8 http://en.wikipedia.org/wiki/Tf–idf

Correspondingly, one can take advantage of the analysis of co-occurrences of subject headings and Dewey classifications in bibliographic records. This is often called statistical mapping. In our case, it is of limited use since our approach accounts for the mapping of many terms that do not co-occur with Dewey classifications

Additionally, there are many instances of bibliographic records with several subject headings and several Dewey classes. In these cases, the connection between the subject heading and the Dewey class is uncertain.

We also see that subject headings from BIBBI (i.e. the bibliographic database elaborated by the Norwegian Library Bureau) could be useful. The database contains ca. 220 000 bibliographic records that have been assigned both a subject heading and a Dewey class (from the latest Norwegian translation, the abridged and modified DDK5). For the subset of BIBBI subject headings that are concurrent with Humord subject headings, we see that they could aid in the suggestions of relevant mapping candidates.

It is still an open question whether we will go for a more general or a more context-dependent mapping between Humord and WebDewey. The challenge in using both the UiO Library's subject index and statistical mapping as a context in the synthetic documents for the Humord-terms is that we may get strong connections to Dewey classes that do not correspond to the correct context for the Humord term. If we go for a general mapping towards all disciplinary perspectives in DDC, this context will be quite useful. If we go for a mapping based on Humord-contexts for each subject heading, this approach will lead to low precision in the results list. This is illustrated with the example on "horses" in the chapter "Intellectual challenges associated with mapping".

We think that the approach can be used for both cases, but it is necessary to reflect over how the use of the subject index and statistics from the catalogue will affect the result.

Use of the subject index from BIBBI will likely lead to a more general mapping that reflects the document collection in a public library.

Application architecture for ccmapper

The prototype ccmapper will, as mentioned, be based on SKOS and SPARQL. In the development of the prototype, we have used the Callimachus framework (callimachusproject.org), which is based on RDFa, SPARQL, XHTML5, CSS3 and JavaScript.

It is currently too early to conclude whether this framework can be used for the final version of ccmapper, which will become operational.

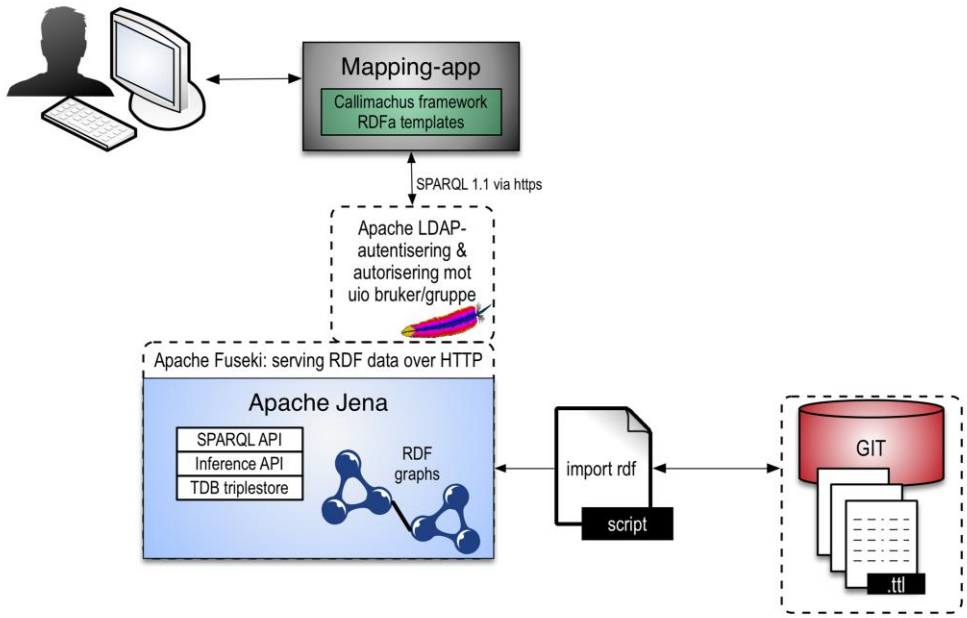

Figure 5: Frame diagram for ccmapper in production

Conclusion

In the present report we have presented the activities which are either finalized or in progress as per ultimo February 2015. It is important to take into account that some of the activities and problems discussed in this report will not be completed within the frame of the pilot project, but will be further prolonged this year.

In the fall semester of 2014, the UiO Library sent a new application to the National Library of Norway for funding to plan and conduct mapping to Norwegian WebDewey. The required funding has been granted. This has given us a unique opportunity to prolonge the project *Methodology for mapping Humord to WebDewey* as the new project *Mapping to Norwegian WebDewey*.

The task of mapping, which will comprise the vocabularies Humord and Realfagstermer, will be performed along the instructions provided in ISO 25964-2 (International Organization for Standardization, 2013). As mentioned earlier, this ISO standard contains general guidelines, and not specific guidelines for mapping to a classification scheme like DDC. Accordingly, it is necessary to develop a theoretical understanding and practical applications which are in accord to our vocabularies, both the source vocabularies Humord and Realfagstermer, as well as the target vocabulary DDC.

The purpose of the *ccmapper* tool is to come up with mapping candidates and thus facilitate the intellectual part of the mapping process. The experiences we gain in our test mapping efforts will be integrated as algorithms in this tool. The functionality of the ccmapper will be further adjusted according to the recommendations for best practices for mapping to Dewey, which will result from the Naples seminar. After these adjustments, the actual mapping of Humord and Realfagstermer to WebDewey can get started. This work will be performed in close collaboration with the Dewey editorial board at the National Library of Norway. We believe that our work will have transfer value for other Norwegian an international library institutions who want to perform similar mappings to Dewey. We will facilitate the sharing of our experiences with other professional environments.

References

Euzenat, J., & Shvaiko, P. (2013). *Ontology matching* (2. utg.), Berlin: Springer.

Hollink, V. Kamps, J., Monz, C, & De Rijke, M. (2004, januar). Monolingual document retrieval for European languages. *Information Retrieval*, *7*(1-2), 33-52. Doi:10.1023/B:INRT.0000009439.19151.4c

International Organization for Standardization. (2009). *Information and documentation: Thesauri and interoperability with other vocabularies: Part 1: Thesauri for information retrieval (ISO 25964-1: 2011).* Geneve: International Organization for Standardization.

International Organization for Standardization. (2013). *Information and documentation: Thesauri and interoperability with other vocabularies: Part 2: Interoperability with other vocabularies. (ISO 25964-2: 2013)*. Geneve: International Organization for Standardization.

Knutsen, U., & Gulbrandsen, A.D. (2014). På randen av mapping. *Bibliotheca Nova,* (4), 36-46.

Kuldvere, V., Lundevall, M., Hegna, K., Konestabo, H. S., Låberg, K. T. , Flatby, E. S., & Greenall, R. (2013, 7. mai). *Realfagstermer og TEKORD: RDF som plattform for sammenlikning og sammenføyning av emnesystemer?: Rapport*. Hentet fra ub.uio.no/om/prosjekter/avsluttet/ realfagstermer-tekord/realfagstermer-og-tekord-rapport.pdf

Kuldvere, V., Flatby, E.S., Heggø, D.M.O, Konestabo, H.S., Lundevall, M., & Låberg, K.T. (2014, 1. juli). *Felles terminologi for klassifikasjon med Dewey: Rapport*. Hentet fra http://urn.nb.no/URN:NBN:no-44610

Salton, G., Wong, A., & Yang, C. S. (1975, november). A vector space model for automatic indexing. *Communications of the ACM,18*(11), 613–620.