EPC Exhibit 137-28.1

February 28, 2014

THE LIBRARY OF CONGRESS

Dewey Section

To: Jonathan Furner, Chair

Decimal Classification Editorial Policy Committee

Cc: Members of the Decimal Classification Editorial Policy Committee

Karl E. Debus-López, Chief, U.S. Programs, Law, and Literature Division

From: Rebecca Green, Assistant Editor

Dewey Decimal Classification

OCLC Online Computer Library Center, Inc.

Via: Michael Panzer, Editor in Chief

Dewey Decimal Classification

OCLC Online Computer Library Center, Inc

Re: Relative Index: Managing synonyms

[This exhibit is the same as Exhibit 136-26, to which Appendix B has been added.]

This exhibit addresses control of synonymy in the Relative Index (RI) and its representation in the MARC Authority Format. In the past, this representation has been governed primarily by cross-reference display considerations. We would like to move, however, to a representation with a solid semantic basis. Since treating the Relative Index as a controlled vocabulary involves certain complexities, we preface our proposed representation by exploring the Dewey editorial rules on synonym control, the MARC Authority Format’s representation of relationships among headings, and the motivation for how the MARC format is currently being used to represent relationships between RIs.

**The Relative Index as Controlled Vocabulary**

In large part, the DDC’s Relative Index is thesaurus-like, having been developed along the lines of a controlled vocabulary[[1]](#footnote-1): Relative Index terms are designed to follow consistent patterns in wording and structure and may be interrelated by equivalence, hierarchical, and associative relationships. Furthermore, the MARC authority format is used to store Relative Index data.

But certain aspects of the Relative Index break out of the controlled-vocabulary mold. For one thing, in keeping with its status as an index for a specific resource, terms in the Relative Index are assigned to DDC classes, not to bibliographic resources. For another, although the Relative Index honors the preferred-term concept for control of synonyms, it implements synonym control somewhat differently from the typical controlled vocabulary.

Standard thesaurus practice involves selecting one term from among a group of (quasi-) synonyms as the preferred/authorized/established term. This preferred term is the only term from the synonym set actually used in indexing bibliographic resources. In the controlled vocabulary, a non-preferred term appears only as lead-in vocabulary to a preferred term: it is not used to index bibliographic resources; its entry in the vocabulary serves only to indicate which authorized term is the preferred term to be used in its stead.[[2]](#footnote-2)

The Relative Index likewise responds to synonymy by designating one term from among a set of synonyms as the preferred term. However, instead of restricting the appearance of non-preferred terms to entries/cross-references leading from non-used/non-preferred terms to used/preferred terms, the import of a preferred term in the Relative Index is to designate which synonymous term is to receive full indexing in circumstances where such a distinction is made.

This distinction between non-established terms and non-preferred terms reflects a greater association between the Relative Index and concept-based thesauri than between the Relative Index and term-based thesauri. (See Appendix A for further elucidation.)

**Synonymy in the Relative Index Editorial Rules**

The Editorial Rules (from the Index Editorial Rules) relevant to synonymy are these:

7.4 **Selection of the synonymous entry receiving full indexing**

7.4.1 **Entries with three or fewer subentries**

When the entry has three or fewer subentries, the subentries and references are given under each form of the entry.

Earthworms 592.64

culture 639.75

Night crawlers 592.64

culture 639.75

Not:

Night crawlers 592.64

*see also* Earthworms

7.4.2 **Entries with four or more subentries**

When the entry has four or more subentries, prefer the vernacular form of the heading given in the schedule or table. If the heading is not one likely to be sought by users, use the sought term if it conforms to other index rules. The form receiving full indexing is referred to from the other forms of the entry (**see 8.2.1.2**).

Aves 598

*see also* Birds

Birds 598

agricultural pests 632.68

air transportation hazard 363.124 12

animal husbandry 636.5

art representation 704.943 28

. . .

8 **References**

8.1 **See reference**

Because each term from which a reference is made always has at least one number (i.e., the number or numbers under which the term is most likely to be classified), see references are not used.

8.2 **See-also reference**

8.2.1 **When to use**

8.2.1.2 To lead to the form of the synonymous entry chosen for full indexing when it will supply at least three additional numbers. However, if the entries are spelled almost alike, then the chosen entry need yield only two extra entries. **See also 7.4.2 for selection of the synonymous entry receiving full indexing**.

Senior citizens 305.26

T1—084 6

*See also* Older persons

**Vocabulary Control in the MARC Authority Format**

According to the *MARC 21 Format for Authority Data*, the MARC Authority Format designates (1) authorized forms of names / subjects / subject subdivisions, (2) other forms of names / subjects / subject subdivisions to be used in references to the authorized forms, and (3) relationships among these forms. Different kinds of authority records have been defined; a key distinction among kinds of records is whether the form in the 1XX field is authorized (008/09 = {a, d, f}) or unauthorized (008/09 = {b, c, g, e}).

Relationships among subject headings are recorded in the 2XX-5XX fields of the MARC Authority Format, as explained in the Definition and Scope statements of the relevant sections:

Cross references lead from an unauthorized heading to an authorized heading (see reference) or from one authorized heading to another (see also reference). Actual cross references are generally not carried explicitly in authority records. Instead, variant form and related headings are carried in authority records in either tracing or reference note fields. Cross reference displays may be generated by combining the contents of a tracing or a reference note field and the 1XX heading field of a record.

More specifically:

Fields 260 [Complex See Reference – Subject] and 360 [Complex See Also Reference – Subject] are used to record information about unauthorized forms of subject headings, different authorized forms of subject headings, and other variants not chosen as the authorized form when the relationships cannot be adequately conveyed by one or more simple cross references.

Fields 400-485, the See From tracing fields, are used to identify unauthorized forms of headings and other variants not chosen as an authorized form.

Fields 500-585, the See Also From tracing fields, are used to identify different authorized forms of headings related to the authorized form in the 1XX field.

**Synonymy among Relative Index Headings as Represented in MARC**

Prior to 2010, DDC data were maintained in a proprietary format. In the conversion of Relative Index data to the MARC format, the following conventions for synonymous terms have been followed:

* In keeping with editorial rule 8.1, no terms have been designated in the Relative Index as unauthorized forms of heading; no 4XX fields have been used.
* The recording of relationships between non-preferred terms and preferred terms has been limited to those circumstances in which see-also references are to be displayed. In such cases (see editorial rules 7.4.1, 7.4.2, and 8.2.1.2), the record for the non-preferred term contains a 260 field for a see-also reference leading to the preferred term; the record for the preferred term contains a 5XX field, with a $w (control subfield) value of “|||a”, meaning that the display of a cross-reference from the preferred term should be suppressed.

Consider, for example, the following two RI records (showing only relevant fields):

001 och00011563 ​

083 04 $a 590 $0 (OCoLC-D)ocd00149863 $2 23 $5 OCoLC-D $9 as=AP $9 ps=PE

150 ## $a Animals ​

550 ## $w |||a $a Fauna $0 (OCoLC-D)och00047216 $2 23 $9 ra=AP $9 rv=PE

001 och00047216 ​

083 04 $a 590 $0 (OCoLC-D)ocd00149863 $2 23 $5 OCoLC-D $9 as=AP $9 ps=PE ​

150 ## $a Fauna ​

260 ## $i see also $a Animals $0 (OCoLC-D)och00011563 $2 23 $9 ra=AP $9 rv=PE ​

which result in the following being printed/displayed in the Relative Index (both in print and in WebDewey):

Animals 590

[plus numerous subheadings]

. . .

Fauna 590

s*ee also* Animals

In short, the current recording of Relative Index relationships between non-preferred terms and preferred terms has been designed completely around the display of see-also references. For example, there are no 2XX–5XX fields in the record for Night crawlers, although editorial rule 7.4.1 implies that Earthworms is the preferred term for the topic and Night crawlers is not. Moreover, we are (mis)using the 260 field, which is intended for complex **see** references, by using it instead to display **see also** references. This decision was made because there are no restrictions on the text placed in 260 $i (that is, because we could make it produce what we needed) and because the similarly defined 360 field, which is intended for complex **see also** references, was put to use for a small set of references of the form “see also entries beginning with [word],” which we wished to keep separate.

**Moving Forward**

In light of our long-term effort to develop the DDC into a system that can support automatic application, it is imperative that our control over relationships among Relative Index terms be governed more by semantic principles than by display exigencies. At the same time, we cannot simply adopt standard usage of the MARC Authority Format for equivalence/synonymy relationships, because the Relative Index’s notion of preferred heading is not the same as MARC’s notion of authorized/established heading. Since the Relative Index *always* gives at least one number for every form of heading that it recognizes, it contains no such thing as an unauthorized/unestablished heading.

This is how we propose to establish synonym control over Relative Index headings in the MARC Authority Format:

* Every Relative Index heading should be represented by a record for that heading in the 1XX field. (This accords with the current situation.)
* Where two or more Relative Index headings have more or less the same meaning, one of them should be designated the preferred heading; 4XX fields should be added to the record of the preferred heading for all other headings among the set of synonyms; a 5XX field for the preferred heading should be added to each of the records for the other headings among the set of synonyms. If a cross-reference should *not* be displayed from a synonym to the preferred heading, the 5XX field should contain “$w |||a.” If, however, a see-also cross-reference should be displayed, no $w subfield will be used.

Note that the proposal calls for a re-interpretation of 4XX fields, which are officially defined as representing non-established/unauthorized forms of heading. We would need to re-interpret them as designating non-preferred forms of heading.

What needs to be done to transition from the data in their current state to this proposed new representation? A number of steps would be needed:

* Identify all RI records with 260 fields. (The 1XX headings in these records are non-preferred terms; the headings in the 260 $a subfield are the corresponding preferred terms.)
* For each record identified in the preceding step, also locate the record referred to in its 260 $a and $0 subfields. For the referring record, convert the 260 field into a 5XX field by changing the MARC tag number and deleting the $i subfield. (The XX portion of the MARC tag is taken from the XX portion of the 1XX tag found in the referred-to record.) For the referred-to record, convert the 5XX field corresponding to the referring heading (i.e., where the 5XX $0 subfield links to the record for the referring heading) into a 4XX field by changing the MARC tag number. All other data in subfields of these fields are retained in converting the fields from 260 to 5XX and from 5XX to 4XX.

For example, the following data in their pre-converted state—

001 och00011563 ​

083 04 $a 590 $0 (OCoLC-D)ocd00149863 $2 23 $5 OCoLC-D $9 as=AP $9 ps=PE

150 ## $a Animals ​

550 ## $w |||a $a Fauna $0 (OCoLC-D)och00047216 $2 23 $9 ra=AP $9 rv=PE

001 och00047216 ​

083 04 $a 590 $0 (OCoLC-D)ocd00149863 $2 23 $5 OCoLC-D $9 as=AP $9 ps=PE ​

150 ## $a Fauna ​

260 ## $i see also $a Animals $0 (OCoLC-D)och00011563 $2 23 $9 ra=AP $9 rv=PE ​

—should become:

001 och00011563 ​

083 04 $a 590 $0 (OCoLC-D)ocd00149863 $2 23 $5 OCoLC-D $9 as=AP $9 ps=PE

150 ## $a Animals ​

450 ## $w |||a $a Fauna $0 (OCoLC-D)och00047216 $2 23 $9 ra=AP $9 rv=PE

001 och00047216 ​

083 04 $a 590 $0 (OCoLC-D)ocd00149863 $2 23 $5 OCoLC-D $9 as=AP $9 ps=PE ​

150 ## $a Fauna ​

550 ## $a Animals $0 (OCoLC-D)och00011563 $2 23 $9 ra=AP $9 rv=PE ​

* Identify all RI records for pairs of synonymous headings where there is no use of 260 fields, that is, where there is no see-also cross-reference display. This is far easier said than done!

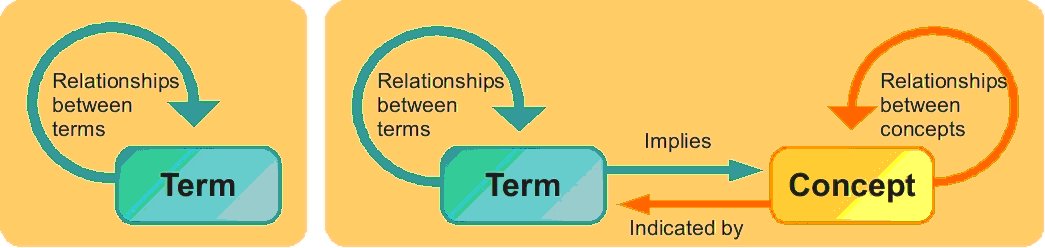
There are multiple subtypes:

* + Unfortunately, in the process of publishing data into the DDC 23 master file, timeout issues sometimes caused the replication of cross-references for headings with a substantial number of cross-references to fail. Ironically, the example used in 7.4.2 is one such case. The record for Aves includes cross-reference fields to Birds for editions 20, 21, and 22 and for the project from which 590 data was published into DDC 23, but there is no cross reference for DDC 23. A comparison of cross-reference data across previous editions, the projects that pushed data into DDC 23, and DDC 23 can be used to identify missing references of this sort (with care taken not to reinsert references that have been intentionally deleted).
  + The task of identifying whether Relative Index cross-reference relationships were equivalence, hierarchical, or associative relationships had not been completed prior to the conversion of Relative Index data to the MARC format. (What had been deemed important previously was only whether a cross-reference should display.) Hence, some Relative Index relationships that should be recognized as equivalence relationships are represented internally as if they were associative relationships. Ironically again, such is the case with Aves and Birds. All RI relationships currently represented as undifferentiated[[3]](#footnote-3) associative relationships should be compared against various online resources such as LCSH (via id.loc.gov) or WordNet to identify synonyms. For example, the LCSH record for Birds gives Aves as a variant term.
  + Where no cross-reference has ever been displayed, it is likely that equivalence relationships have not been recorded. Such is the case (as previously mentioned) for the equivalence relationship between Earthworms and Night crawlers. Some of these missing relationships could be identified on the basis of examining data patterns that match the situation in editorial rule 7.4.1, that is, where two headings and their corresponding subheadings index the exact same classes. Other missing relationships might be identified on the basis of our prior analysis of Relative Index data as presented in EPC Exhibits 129-36.1 and 130-37.1 or extensions thereof.

The processes outlined above can suggest possible equivalence relationships, some with greater certainty than others, but many of them will require human review to verify if an equivalence relationship actually exists between two headings. Contrariwise, some missing equivalence relationships will probably go undiscovered by all of these processes and will need to be discovered and coded on an ad-hoc basis.

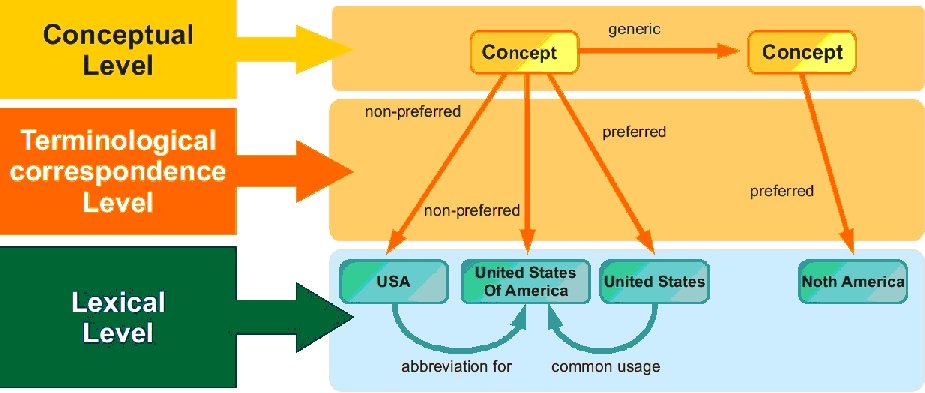
**Appendix A**

Excerpted from: J.-A. Pastor-Sanchez, F. J. Martínez Mendez and J. V. Rodríguez-Muñoz. “Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives.” *informationresearch* 14/4 (December 2009). Available: http://informationr.net/ir/14-4/paper422.html



**Figure 1: Term-based thesaurus compared with concept-based thesaurus.**

In a term-based thesaurus the semantic relationships are established between the terms themselves. Meanwhile, in a concept-based thesaurus, certain semantic relationships are established between concepts and others between terms. The relationships between concepts generally include those which develop the hierarchical and associated structures, while the latter denote the relationships between terms associated with one single concept (abbreviations, common usage, technical language, etc.). A term-based thesaurus is thus structured at one single lexical-terminological level, whereas a concept-based thesaurus has a three-level structure: (a) conceptual level, where concepts are identified and their interrelationships established; (b) terminological correspondence level, where terms are associated (preferred or non-preferred) to their respective concepts and (c) lexical level where lexical relationships are defined to interconnect terms.



**Figure 2: organizational levels of a concept-based thesaurus.**

**Appendix B**

We received the following comments on this exhibit in its previous manifestation as EPC Exhibit 136-26 from Elise Conradi, Nasjonalbiblioteket / National Library of Norway:

30 April 2013

Hi all,

I was very pleased to see the proposals outlined in EPC 136-26! For some time, we have been concerned about the management of synonyms in WebDewey. It is good to see that you have had the same concerns, and that you propose to link preferred and non-preferred synonyms in the Relative Index regardless of the display of see-also references. I understand that there are many steps to take to identify synonyms in the current data. If there is anything we can do at the National Library of Norway to help in this endeavor, please let us know. For example, you write on page 8 that it is “likely that equivalence relationships have not been recorded” when there hasn’t been a need to display cross-references. For some time, we have been recording all added Norwegian synonyms in an external file, noting preferred vs non-preferred terms, plus which have been linked in the 260-field and which haven’t. So if you would like to test out any of the proposal on our data, we’d be happy to oblige!

Btw, I’ve noticed that none of the links from preferred terms to non-preferred terms (in the 550 field) are available in the data in our translation software. We do, however, have the 260-field links – so hopefully these can be used to generate corresponding 550-field links (before they are all converted to 450 and 550-fields).

There are still a couple of major (?) problem regarding synonym management in WebDewey that have not been addressed by the EPC-exhibit:

1. How can we link synonyms between Relative Index terms for which there is no interdisciplinary placement of the topic? Here is a short list of examples of English terms without interdisciplinary placements, for which we would like to use more than one Norwegian terms:

* Artificial insemination [Kunstig befruktning (preferred) + Kunstig inseminasjon]
* Conversation [Konversasjon (preferred) + Samtale]
* Euthanasia [Dødshjelp (preferred) + Eutanasi + Barmhjertighetsdrap]
* Exorcism [Eksorsisme (preferred) + Djevelutdrivelse]
* Zodiac [Dyrekretsen (preferred) + Stjerentegn + Zodiaken]
* Inscriptions [Innskrifter (preferred) + Inskripsjoner]

Juli mentioned while she was here in Oslo that we should send each of these cases to you, so that you could decide whether to find interdisciplinary placements for them. We’re wondering if perhaps it would be possible to reconsider the rule that each relative index term must be associated with a class? The latter would allow for the creation of records for topics that don’t have interdisciplinary placements, but that do have more than one term representing it.

1. We are very concerned about the display of preferred vs non-preferred terms in WebDewey. We fear that it is unduly complicated for end-users (and even sometimes for editors!) to figure out which of two terms is to be preferred when, for example, building new numbers and adding user terms. In light of the role Relative Index terms are beginning to play in document retrieval (for example in WebDewey Search), we feel that it is imperative that number-builders clearly see which of two terms is preferred. Perhaps one of the following options could be considered?

* Non-preferred terms are not displayed in the class record or in the NBE. They continue to be displayed in results lists for search and browse in WebDewey, following today’s practice\*.
* Non-preferred terms are displayed in class records and in the NBE as “belonging” to preferred-terms, for example indented under the preferred terms or with a typographical distinction. They continue to be displayed in results lists for search and browse in WebDewey, following today’s practice.

(\*A propos today’s practice: we find the use of see-also notes to denote equivalence relationships to be confusing, considering that they are used entirely differently for thesaurus relationships. I see you address this in the exhibit, but I remain unclear as to how these types of references will be displayed in result lists in the future.)

Let me know if you have any questions regarding these concerns and ideas!

Kind regards,

Elise

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Elise Conradi

Norsk webDewey prosjektleder

Nasjonalbiblioteket - Avdeling for Tilvekst og kunnskapsorganisering

[elise.conradi@nb.no](mailto:elise.conradi@nb.no)

+47 23 27 62 44

We sent the following reply:

Hi Elise,

This is an unofficial response / mostly my own thinking, but I wanted you to know that your email had not gone “plunk” into the pond in its travel across the Atlantic without creating some waves. (We’re all into “ripple effects,” as you know.)

1. We, in fact, do have records for Relative Index terms for which no interdisciplinary works number has been designated (i.e., with no 083 fields). A very small number of them carry equivalence relationships (considerably more of them carry hierarchical relationships). For example:

LDR nz###n##

001 och00008847

003 OCoLC-D

005 20110303130633.0

008 100206|||a|z||||||##########||#a||#####d

040 ##  $a OCoLC-D $b eng $c OCoLC-D $d OCoLC-D $f ddcri

150 ##  $a Agricultural laborers

260 ##  $i see also $a Agricultural workers $0 (OCoLC-D)och00008900 $2 22 $9 ra=AP $9 rv=PE $9 uts=2011-03-03T13:06:33Z-Rebecca Green

260 ##  $i see also $a Agricultural workers $0 (OCoLC-D)och00008900 $2 23 $9 ra=AP $9 rv=PE $9 uts=2011-03-03T13:06:33Z-Rebecca Green

(Juli points out that agricultural workers and agricultural laborers are not equivalent terms in the schedules [see 305.563]. So take the example with a huge grain of salt.)

I suspect there may be a distribution issue here, as well as a display issue. But I don’t think there’s an inherent issue in recording the relationships in the first place.

2. How to make preferred terms more visible was outside the scope of EPC 136-26, but I think you provide cogent reasons why they need to be more visible and offer some valid ways of doing it. I’m sure we will want to consider this further. Thanks for raising the issue.

[*Further elaboration:* We agree that we need better ways of displaying which term among sets of equivalent terms is the preferred term. Elise offers some suggestions that could be implemented in WebDewey based on the coding suggested here. Do we also need some way of making preferred RIs more obvious to editors/translators? The transformations recommended in this exhibit will cause the reference fields in DDC authority records to appear more like those in, e.g., LCSH, records. The likelihood of those transformations making preferred RIs “obvious” is reasonably high.]

3. I wonder if DDC’s use of “see also” is cast in stone. Since our distinction is between preferred and non-preferred, and not between authorized and unauthorized, it could make sense to adopt different wording. Something else for future thought.

[*Further elaboration:* What if we used the word “prefer” instead of the phrase “see also”?]

Rebecca

1. In reality, the Relative Index is what it says it is—an index—and follows standards developed for indexes (e.g., *ISO 999: Information and Documentation—Guidelines for the Content, Organization and Presentation of Indexes*). The general principles governing synonym control in indexes are similar to those used in thesauri, which are more widely understood. [↑](#footnote-ref-1)
2. Standard index practice also involves selecting one term from among a group of (quasi-) synonyms as the preferred term. Cross-references lead from headings not used in the index to terms used in the index. [↑](#footnote-ref-2)
3. Some relationships are distinguished by the value of the first position of the control subfield—$w g for broader term, $w h for narrower term. Where this is the case, the cross-reference relationship need not be processed to determine if an equivalence relationship exists. [↑](#footnote-ref-3)