Submitted on: 22.10.2017

2016 Satellite meeting - *Subject Access: Unlimited Opportunities*
11 – 12 August 2016
State Library of Ohio, Columbus, Ohio, USA

Crowdsourcing the Dewey Decimal Classification: When Users Become Contributors

Elise Conradi
Acquisition and Bibliographic Services, National Library of Norway, Oslo, Norway.
elise.conradi@nb.no

Rebecca Green
Dewey Decimal Classification, OCLC, Washington, DC, USA.
greenre@oclc.org

Alex Kyrios
Dewey Decimal Classification, OCLC, Washington, DC, USA.
kyriosa@oclc.org

Abstract:

Over half of all Dewey Decimal Classification (DDC) numbers in bibliographic records are built numbers—numbers generated by following instructions to add notational components from various parts of the DDC to base numbers. WebDewey includes a number building tool that assists users in building Dewey numbers and permits them to contribute those numbers and their corresponding user terms for review and approval; vetted numbers are returned to WebDewey with accompanying Relative Index terms so other users can access and use the contributed built numbers.

This paper describes the initial behavior of users from the English and Norwegian WebDewey communities in contributing Dewey numbers and access terms. Phenomena examined include, for example, the complexity of the numbers that users build, the frequency with which users accept default terms, and the kinds of changes that users make in the contributed terms.

Keywords: Dewey Decimal Classification, WebDewey, user contribution, synthesized numbers, number building tool.

Complex subjects—subjects involving two or more concepts and the relationships that bind them—are a fact of the modern bibliographic world. Classification systems often support complex subjects through the combination of notational components that express elemental concepts and their interrelationship. The Dewey Decimal Classification (DDC) system, in

particular, supports complex subjects through the number-building process. Built numbers rely on explicit add instructions or on the permissibility of adding standard subdivision notation to any valid Dewey number, unless specifically prohibited. Entries for some built numbers are given in the schedules and tables so that relevant notes for the classes can be presented. Many other built numbers are available solely as the target of Relative Index headings. (Relative Index headings express subjects and provide access to DDC classes.)

As its name implies, WebDewey is a web-based application through which users (typically catalogers/classifiers) can access information about the DDC. One of its features, the number building tool, guides users through the process of building Dewey numbers. After saving a built number, the user is given the opportunity to create one or more user terms to facilitate future access to the newly built number. The WebDewey number building tool (Beall 2016, 2012a, 2012b) was first introduced into the English WebDewey in November 2012, and its functionality has been updated periodically since. The number building tool uses instructions coded in MARC records for classification numbers and spans. These instructions facilitate building from add tables and designated base numbers. In their absence, it defaults to Table 1 Standard Subdivisions, which can be applied throughout the schedules unless otherwise specified.

Human intervention is still sometimes required when using the number building tool, especially in instances where standard subdivisions are displaced to more than one zero; while the number building tool can prevent a user from building invalid numbers in some cases, it cannot do so perfectly, and some users have contributed invalid numbers. Having members of the editorial team review contributed numbers helps screen out invalid numbers, and analysis of invalid numbers that make it through the number building tool can help alert editors to common errors, or unclear instructions in the schedules.

By default, the numbers and associated headings created by users are available in WebDewey only to their creators or other members of the same institution. However, users can also contribute numbers and user terms for general use, in which case the numbers and headings are vetted, modified as necessary to maintain consistency across Relative Index headings, and fed back into WebDewey. In this manner, the repository of valid built DDC numbers expands through crowdsourcing.

This paper describes the initial behavior of users from the English and Norwegian WebDewey communities in contributing Dewey numbers and access terms. How complex are the numbers that users are building? How often do users contribute terms just by accepting the default terms initially supplied for them? We also compare and contrast the contributions of Norwegian and English WebDewey users. Lastly, we explore the impact of crowdsourcing the DDC: what patterns emerge in the contributions that users are making to the classification system?

Number Building in the DDC

The DDC English database consists of numbers that are assignable as complete Dewey numbers, as well as number components that can be used in building assignable numbers. At present, 40,550 assignable numbers reside in the database, of which 22,345 (55%) are integral / cannot be broken down into parts, while another 18,205 (45%) are built/synthesized numbers. The database also includes 12,191 number components—building blocks from Tables 1-6 and internal add tables—that can be added in specific contexts to express such topics as language, geographic area, ethnic/national group, genre, etc.

The set of assignable numbers includes 902 numbers with add instructions (including 19 of the assignable numbers that are built numbers with add instructions of their own). The set of records for number components supplies another 268 add instructions. Some add instructions give directions for appending notation from number component records (e.g., "Add to base number 410.18 the numbers following 4—018 in Table 4, e.g., cognitive linguistics 410.1835"), while other add instructions give directions for appending notation derived from assignable numbers (e.g., "Add to base number 156.2 the numbers following 152 in 152.1–152.8, e.g., comparative reaction-time studies 156.283"). As just seen, the notational range accessed by some add instructions is relatively narrow, while for other add instructions the entire range of assignable numbers can be used (e.g., "Add to base number 153.94 notation 001–999, e.g., tests for musical ability 153.9478"). Many add instructions are simple, directing the user to add a single number component, but some add instructions are complex (e.g., "Add to base number 571.99 the numbers following 579 in 579.2–579.5 for the organism only, e.g., protozoan diseases 571.994; then add 1 and to the result add further the numbers following 571 in 571.1–571.2, e.g., protozoan diseases in animals 571.99411, protozoan diseases in plants 571.99412"). It is safe to say that the potential set of built numbers far exceeds the set of numbers provided explicitly within the system.

Moreover, built numbers are in common use in actual assignment of Dewey numbers to bibliographic resources. An analysis of DDC numbers from the current (23rd) edition in WorldCat records reveals that of all the Dewey numbers used, only 12% already exist in the DDC database, with 7% not being built numbers, and 5% being built numbers; a whopping 88% are numbers that are not given explicitly in the database and therefore presumably are built numbers. When the computations are weighted by the number of records to which the Dewey numbers have been assigned and the number of holdings to which the Dewey numbers have been assigned, the proportions are not so lopsided. When taking records into account 45% are non-built existing numbers, 31% are built existing numbers, and 25% are built and not previously existing numbers; when taking holdings into account the rounded proportions are 42%, 39%, and 20%, respectively. (The decrease in built numbers not in the database from 88% to 25% and 20% is a good thing; it means that a large portion of the built numbers needed for actual bibliographic works have been provided in the Dewey database.)

The Number Building Tool

The contribution of user-generated built numbers and Relative Index terms is encouraged so users of WebDewey will not need to build numbers repetitively. In the Norwegian context, the numbers and terms can also be incorporated into WebDewey Search (Conradi et al. 2016) and other end-user applications based on Norwegian Dewey data. After Norwegian contributed numbers are vetted, they are incorporated into Norwegian WebDewey as "standard" (or national) built numbers, the numbers are displayed in their respective hierarchy in WebDewey Search, and their Relative Index terms are indexed and provide an access point to the class (and from there, provide access to documents with the associated class notation).

For an example of how the number building tool functions, suppose a WebDewey user wants to build a number for the psychology of smell. The user can build the number 155.91166 for the topic, starting at the entry for 155.911 Influence of sensory stimuli, which contains the add instruction, "Add to base number 155.911 the numbers following 152.1 in 152.14–152.18." Clicking Start begins the number-building process. The engine immediately takes the user to a display of the hierarchy under 152.1 Sensory perception, where the user can drill down to

152.166 Perception of smells. Clicking Add at this number results in the desired built number 155.91166, with the number building tool identifying the numbers following 152.1 in 152.166 (i.e., 66) and appending these numbers to the previously established number. Where add instructions allow, the number building tool is capable of adding successive number components to build Dewey numbers of great complexity.

Figure 1 illustrates this example, showing a WebDewey screen with 155.91166 built. From here, a user can click Save to proceed, or continue to add more components. The figure shows a standard subdivision, T1—0721 Research methods, which could be added to the current number to make 155.911660721, representing research methods in smell psychology.

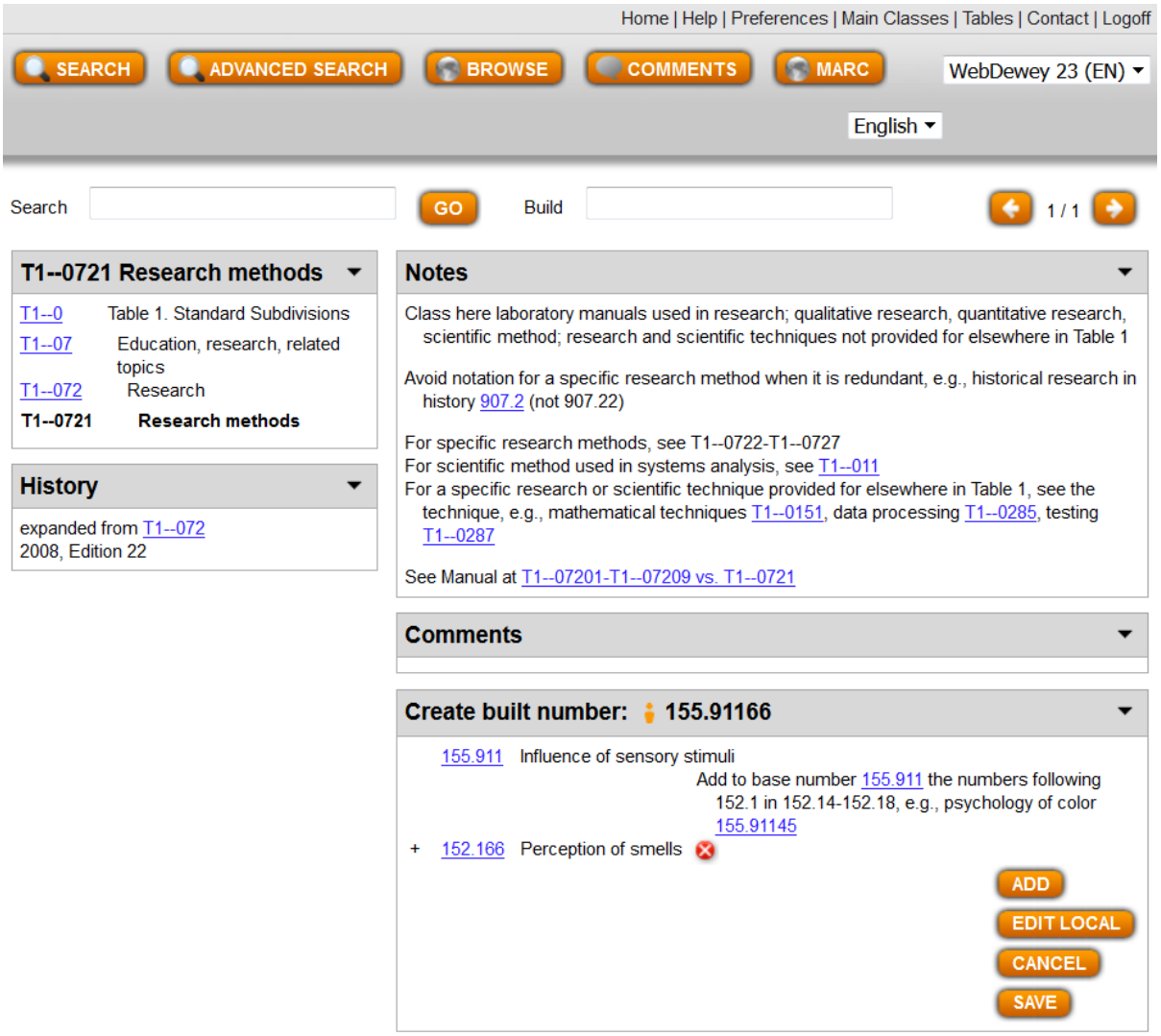

Figure 1: A number built for the psychology of smell

After a user decides to save a number, the number building tool enables the input of user terms for it. One purpose of user terms is to function like the captions given for numbers in the schedules, simply to name what the number represents. A second purpose of user terms is to provide access to the newly built number. The number building tool automatically constructs a default user term, drawing from the captions of the individual components that make up the built number.

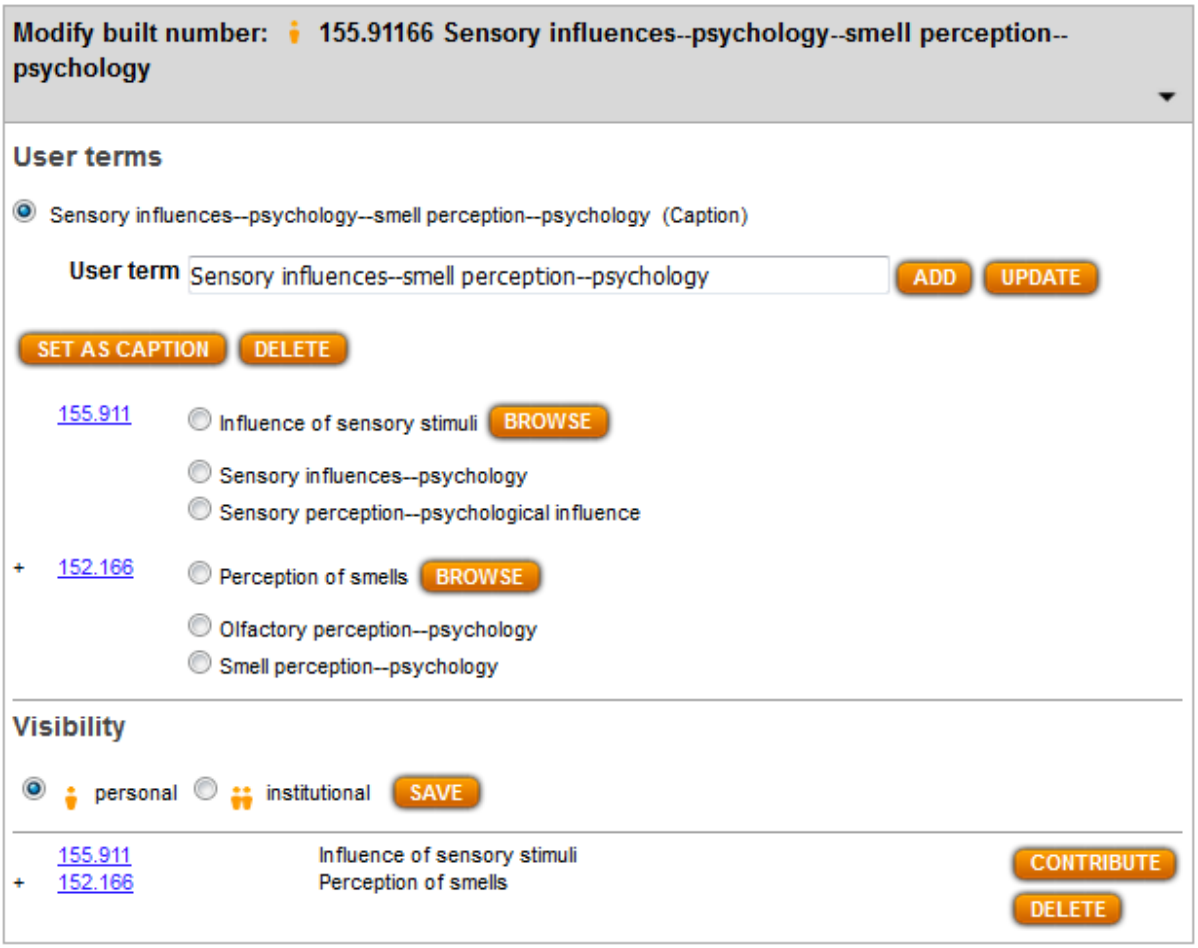

Figure 2: User terms for 155.91166

Figure 2 shows the process of providing one or more user terms for the newly built number. The number building tool provides its default term, Influence of sensory stimuli—Perception of smells, based on the captions at 155.911 and 152.166. The number building tool also presents Relative Index term options for the components. The user might then create, instead of or in addition to the default term, the user term "Sensory influences—smell perception—psychology."

The number building tool's default strategy for generating terms makes some number of assumptions:

- The best Relative Index term for a class is one that approximates the whole of the class.
- Captions are more likely to reflect the whole of a class at a number than any given Relative Index heading at that number.
- In general, each component of a built number should be reflected in a Relative Index term for the number; that is, the number of components in a built number and the number of subfields in a heading should match.
- The order of subfields in a Relative Index heading should mirror the order of the corresponding components in the number.

None of these assumptions is universally valid:

- Some classes are too broad or vague (e.g., 070 Documentary media, educational media, news media; journalism; publishing; 781.53 Music in specific settings) to be summed up in a single Relative Index term.
    a. Consequently, the captions of such classes are not good choices to be converted into Relative Index terms.
    b. For classes where standard subdivisions are added for all the topics mentioned in the caption together or for one or more individual topics, "the best Relative Index term" often ends up being several equally good and more specific relative Index terms.
- Some components of built numbers are best reflected by Relative Index terms that express context (e.g., discipline, time), typically captured in a separate subfield. In such cases (e.g., 123.3 Chance, indexed by $a Chance $x philosophy; 940.4275 Eastern front, indexed by $a Eastern front $x World War I $y 1916) the cardinality of the component/subfield relationship is not 1:1.
- Occasionally, multiple components of a built number may be best expressed by a single word or phrase (e.g., 629.4092 $a Astronautical engineers, rather than $a Astronautics $x biography). In such cases, the cardinality of the component/subfield relationship is again not 1:1.
- The Dewey editorial rules for Relative Index terms bar recapitulation of the schedules, except in certain cases (standard subdivisions, historical periods, approach/form/genre). This means that, in general, the order of subfields in a Relative Index heading should **not** mirror the order of the corresponding components in a built number.

As a result, it is often the case that the default terms generated by the number-building tool do not make good Relative Index terms.

The Norwegian Use Case

In conjunction with the launch of the Norwegian WebDewey in September 2015, extensive training was given to classifiers nationwide during which they were encouraged to contribute built numbers to the editorial team at the National Library for vetting. It was argued that nationally published built numbers would not only make classification work more efficient in Norway, but that these would also be searchable in the end-user solution WebDewey Search. WebDewey Search is an open website that provides subject access to Biblioteksøk, the Norwegian union catalog, through the navigation of Dewey hierarchies and through word-based queries to Relative Index terms. Classifiers were also advised in training to choose either the default terms or terms from the displayed lists that most closely matched the subject of the document being classified. If the subject was more specific than the listed terms, classifiers were told that they were free to add their own terms. The main message delivered was that they needn't spend a lot of time worrying about which terms to add; the editorial team would ensure the publication of preferred terms that fit the correct pattern.

The Norwegian editorial team had voiced concern that trying to teach users how to find correct terms might serve as an unnecessary deterrent to the contribution of numbers, for several reasons. First of all, finding the correct Relative Index pattern is often quite difficult. There is no set citation order for terms associated with number components, and in some cases, there is more than one pattern to follow. Secondly, there is little synonym control in the Relative Index. Editors were worried that users might get discouraged if they spent a lot of time looking for

the correct Relative Index pattern(s), only to find that their contributions were modified in the vetting process to a term they would have had no way of knowing was preferred.

The vetting process has two parts. The first part concerns whether the built number is correct. Does the built number already exist? If applicable, were the correct number of zeros added? Since only those topics that approximate the whole of a number are subject to number building, was it appropriate to build the number? This stage has revealed itself to play an important role in the quality assurance of DDC classification in general in Norway: not only are incorrectly built numbers rejected, but numbers that seem far-fetched are double-checked against newly classified documents insofar as this is possible. If there is a more suitable number for the document, classifiers are notified.

The second part of the vetting process is a far more time-consuming process. First, editors check to see if a pattern exists in built numbers in the relevant hierarchy or in parallel hierarchies. Then, terms are modified as needed. This can entail changing a user term to a synonymous term that is designated as preferred, or adding other terms that editors think may be helpful to other users. When new user terms are suggested, editors first check to see how the terms relate to the class. If they are synonymous with existing terms, they are added to the Relative Index at the appropriate number component. If they represent new topics, they are sent to the standard English-language Dewey editorial team for vetting.

At the time of writing, Norwegian classifiers have contributed 4026 built numbers, of which 2604 (65%) have been vetted; 2232 (86% of those vetted) have been published in the Norwegian WebDewey, and 372 (14% of those vetted) have been rejected. Classifiers at either the National Library or Biblioteksentralen have contributed 2387 of the numbers, while 381 have come from classifiers at the University Library of Oslo. A project team at the University Library of Oslo responsible for mapping two locally controlled vocabularies to the Norwegian WebDewey have sent in 941 built numbers needed to match their terminologies. The remaining 317 have been sent in by a large number of other libraries, both public and private.

Based on the experiences garnered so far, the Norwegian editorial team has suggested a number of modifications to the user terms section of the number building tool, and to areas in WebDewey and WebDewey Search most affected by user contributions.

Figure 3 shows suggested amendments to the user terms screen in the number building tool. The main difference is in the display of captions and in the treatment of default terms. Captions are displayed in bold font and can only be chosen as user terms if no Relative Index terms are associated with the number component. The default term for the built number is based on the alphabetically first Relative Index term in each component. The reasoning behind the change is that captions do not always coincide with Relative Index terms and should therefore not be used as Relative Index terms. By making a clear distinction between the types of terms available, it is argued that users will more easily choose from the Relative Index terms that most closely reflect the content of the classified material.

Another suggested change to the screen involves the text box. In the current number building tool, the text box is pre-filled with the default term and placed near the top of the screen. In the suggested modification, it is left empty and moved underneath the existing terms associated with each component, and accompanied by the following text: “Choose from the above terms or suggest own terms.”

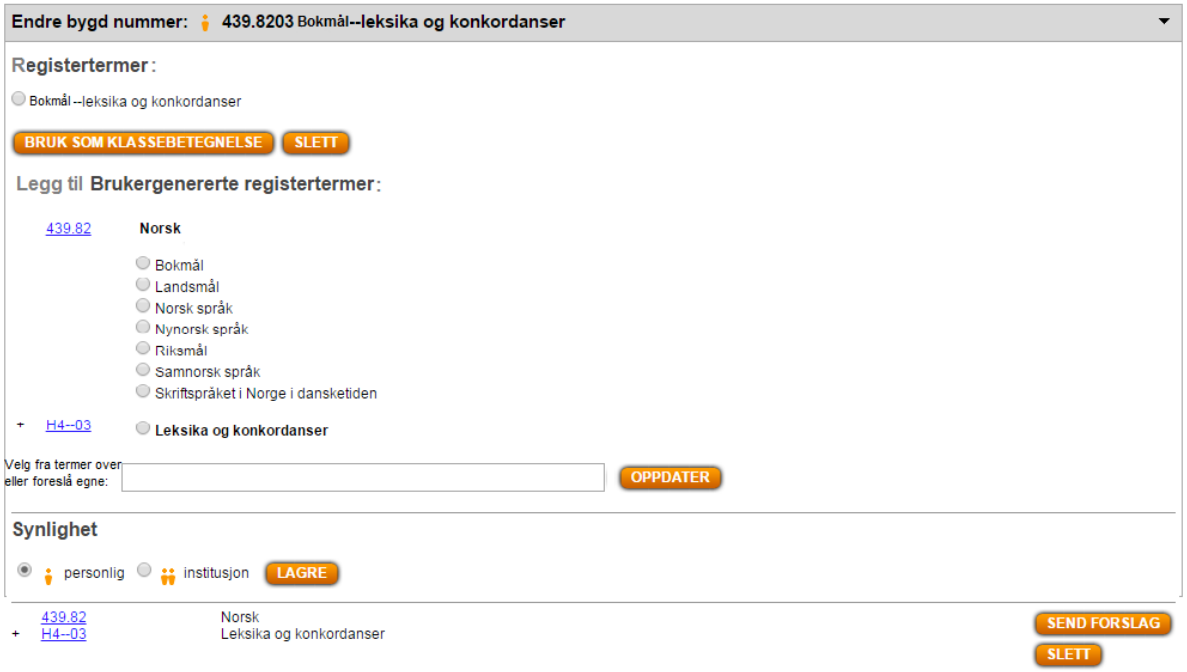

Figure 3: Proposed modifications to user term contribution screen

Since built numbers are included as access points to classified materials in WebDewey Search, classifiers have been increasingly encouraged to choose user terms that most closely match the topic of the document they are classifying. Ideally, vetters would supplement contributed terms with all relevant preferred Relative Index terms for each built number it publishes. This has proved to be quite time-consuming, however. As compensation, the Norwegian editors have suggested another new feature in WebDewey in which users are able to contribute user terms to existing built classes.

Contributed Numbers and User Terms

In general, users contributed numbers with only 1 or 2 build steps, with a fair number of them consisting of only a base number and either Table 1 (standard subdivisions) or Table 2 (area) notation. Table 1 summarizes some of the differences in contributed numbers between users of English WebDewey and Norwegian WebDewey. With its larger number of built numbers with more than 3 components and its smaller number of built numbers consisting of only a base number and Table 1 notation, Norwegian WebDewey apparently sees more sophisticated use of the number building tool.

	English WebDewey	Norwegian WebDewey
Built numbers consisting of base number and Table 1 notation	38%	13%
Built numbers consisting of base number and Table 2 notation	17%	16%
Built numbers using more than 2 build steps / 3 components	3%	11%

Table 1: Character of contributed numbers

Figure 4 displays the interaction between main classes and contributed numbers across English and Norwegian WebDewey combined.

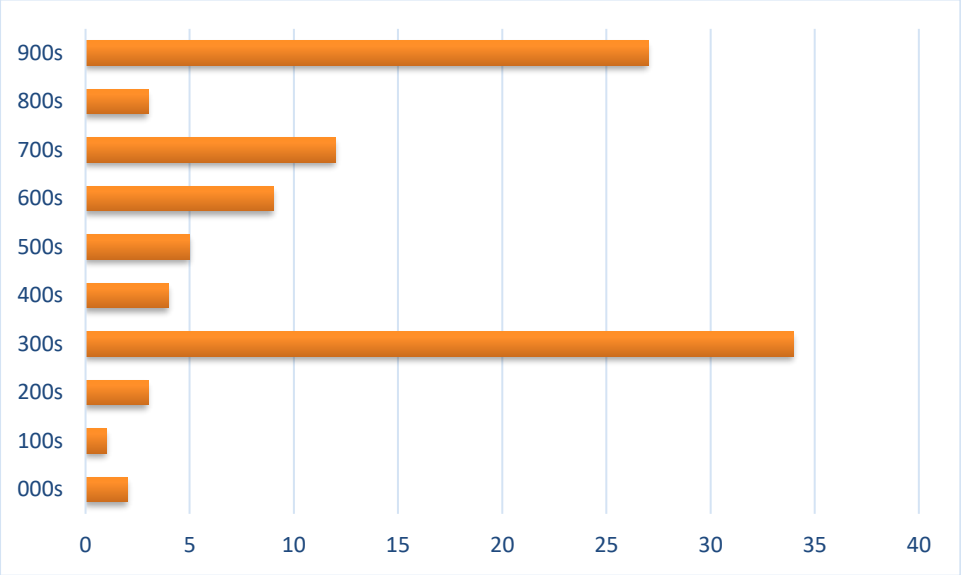

Figure 4: Percentage of contributed numbers per main class

With respect to contributed terms, English WebDewey and Norwegian WebDewey see very similar use, as evident in Table 2.

	English WebDewey	Norwegian WebDewey
Single term per number contributed	90%	88%
Two terms per number contributed	8%	10%
Three or more (up to six) terms per number contributed	2%	2%

Table 2: Number of terms contributed

To get a closer look at user behavior, we examined 100 contributed numbers each from Norwegian and English WebDeweys. Some of the built numbers (16 from Norwegian and 24 from English) were rejected in review. The remainder of the contributed term analysis concerns the 84 correctly built Norwegian numbers and 76 correctly built English numbers.

More often than not, users accept the default term as is—over 75% of the time in Norwegian WebDewey (in accordance with provided instructions), over 50% of the time in English WebDewey.

Vetting and analysis of user-contributed terms require identification of patterns for Relative Index terms appropriate to the corresponding built numbers. On average, 3 changes are needed to convert a user term into a term consistent with the relevant Relative Index pattern, where changes may be any of the following (changes are noted occurring with low, medium, or high frequency):

- Replacing a subfield in the user term with
  - a Relative Index term (medium)
  - the user's own word (low)

    - a semantically related word (whether a synonym [low], a broader term [low], or a narrower term [medium])
    - a shortened version of a term (often a word instead of a multiword expression; medium)
- Capitalizing one or more words in the user term (high)
- Introducing a parenthetical qualifier (medium)
- Switching the order of subfields in the user term (medium)
- Changing the MARC subfield code for a user term (high)
- Deleting a term (high)
- Adding a term (medium)

But sometimes user terms do match Relative Index patterns. In the sample of Norwegian WebDewey user contributions, 10 (12%) of contributed terms matched Relative Index patterns; on 8 of those occasions, the default term accepted by the user matched the Relative Index pattern for the built number as is. In the sample of English WebDewey user contributions, 11 (14%) of contributed terms matched Relative Index patterns; on 9 of those occasions, users had changed the default term. When user terms do not match Relative Index patterns, they are modified during the vetting process before being made available to other users. However, all headings contributed by users continue to be available to them and other members of their institution.

English WebDewey users were much more likely than Norwegian WebDewey users to change a default term before submitting their contribution. The most common kinds of changes made were deleting a subfield in the user term or replacing a subfield with either the user's own term or a Relative Index term option.

Crowdsourcing as a Development Strategy

User contribution is proving successful, especially in Norwegian WebDewey, where contributed numbers and terms enhance the discovery potential of WebDewey Search. But crowdsourcing has limitations as a development strategy for the DDC. The power of Dewey numbers lies in the numbers having definable meaning and in well-defined topics having a single best expression as a Dewey number. The strength of the Relative Index similarly lies in the degree to which it is a controlled vocabulary. Changes made in the vetting process show that user contributions of numbers and terms require intervention if they are not to dilute those strengths.

Acknowledgments

DDC, Dewey, Dewey Decimal Classification, and WebDewey are registered trademarks of OCLC Online Computer Library Center, Inc.

References

Beall, J. "WebDewey Number Building and Two Add Tables in History," *025.431: The Dewey blog* (blog), March 24, 2016, http://ddc.typepad.com/025431/2016/03/webdewey-number-building-and-two-add-tables-in-history.html. [References to many other Dewey blog posts on the number building tool are given at the beginning and end of the post.]

Beall, J. "WebDewey Number Building Tool: User Terms and Predominant Patterns in Relative Index," *025.431: The Dewey blog* (blog), November 29, 2012,

http://ddc.typepad.com/025431/2012/11/webdewey-number-building-tool-user-terms-and-predominant-patterns-in-relative-index-.html.

Beall, J. "WebDewey Number Building Tool: Add Instructions and Base Numbers," *025.431: The Dewey blog* (blog), November 13, 2012, http://ddc.typepad.com/025431/2012/11/webdewey-number-building-tool-add-instructions-and-base-numbers.html.

Aagaard, H., Conradi, E., & Mengel, T. "Leveraging the Dewey Decimal Classification for online subject access: Three use cases of *WebDewey Search*." Paper presented at Subject Access: Unlimited Opportunities, Columbus, Ohio, August 11-12, 2016.