

Diane Vizine-Goetz
OCLC, Online Computer Library Center, Inc.

Julianne Beall
Library of Congress

Using literary warrant to define a version of the DDC for automated classification services

This is a pre-print version of a paper presented at the Eighth International ISKO Conference, 13-16 July 2004, London, UK. Please cite the published version; a suggested citation appears below. The pre-print contains supplemental materials that are not available in the published version. See Appendix 1.

Abstract: This paper presents the results of an exploratory study to determine literary warrant for topics in electronic resources. The classification numbers in Abridged Edition 14 were used as a starting point. Using the principles of abridgment and expansion in Dewey¹, a version of the DDC is defined that accommodates the topics found on three diverse web sites that use Dewey: BUBL, Canadian Information By Subject, and KidsClick! The resulting classes are used to create a database for automated classification of web resources.

Copyright 2004 OCLC Online Computer Library, Inc.
6565 Frantz Road, Dublin, Ohio 43017-3395 USA
<http://www.oclc.org/>

Reproduction of substantial portions of this publication must contain the OCLC copyright notice.

Suggested citation:

Vizine-Goetz, Diane, and Julianne Beall. 2004. "Using Literary Warrant To Define A Version Of The DDC For Automated Classification Services." In Knowledge Organization and the Global Information Society; Proceedings of the Eighth International ISKO Conference, 13-16 July 2004, London, UK, ed. Ia C. McIlwaine. (Vol. 9 in the Advances in Knowledge Organization series; ISBN 3-89913-357-9.) Würzburg (Germany): Ergon Verlag.

1. Introduction

Interest in automatic and semi-automatic assignment of Dewey Decimal Classification (DDC) numbers to electronic documents has remained steady over the past 5 years. Tools to help users assign DDC class numbers were introduced in the Cooperative Online Resource Catalog (CORC) project when the system came online in 1999. These tools, which include automatic generation of DDC numbers for web resources, authority control for DDC numbers, and access to a web version of the Dewey Decimal Classification are now available in the OCLC Connexion Service².

OCLC researchers have reported on how the DDC has been adapted to accommodate automated classification services. These adaptations have focused on optimizing the content of the classification records used for automatic classification. For example, various notes fields, which do not describe a given class, e.g., class-elsewhere-notes, see references, etc., but which describe what is found in other classes, are excluded from the automatic classification database (Shafer, Thompson and Tkac, 1997). The content of classification records is further enriched through the addition of supplemental terminology (Vizine-Goetz, 2001). Additionally, editorial updating of DDC captions in the summaries and in the schedules has improved the expressiveness of many classification records (Mitchell, 2003).

A version of DDC 22 that incorporates many of the improvements described above is being used in the e-Prints UK project. In this project, machine classification services are being developed to classify e-prints and metadata harvested from e-print repositories at UK educational institutions. In a project such as this, it is not clear what edition of Dewey, DDC 22 or Abridged 14, should be used in the classification service. Although the total number of harvested documents is not very large, about 11,600, because the collection contains materials from higher education institutions across a range of scientific and technical disciplines, it was decided that DDC 22 should be used. The DDC classification service is expected to perform adequately for the purposes of the e-Prints project; however, there may be opportunities to define alternate versions of the DDC for machine classification of collections of different size and scope.

This paper presents the results of an exploratory study to determine literary warrant for topics in electronic resources. The study starts with Abridged Edition 14, and using the principles of abridgment and expansion in Dewey, defines an alternate version of the DDC to be used as the source database for automated classification of web resources.

2. Literary Warrant

According to the Dewey Decimal Classification Glossary³, literary warrant is “justification for the development of a class or the explicit inclusion of a topic in the schedules, tables, or Relative Index, based on the existence of a body of literature on the topic.” Addressing how the DDC accommodates translations, Beall (2003) explains how the basis for literary warrant can vary depending on the needs of a given language edition. She explains that the Dewey editors, responsible for the English language editions, rely primarily on the topics found in OCLC WorldCat (the OCLC Online Union Catalog), the Library of Congress catalog, and input from countries that use the English-language edition to determine literary warrant. Translation teams, however, consult the catalogs of major libraries in their linguistic area. For example, the team preparing the German translation uses the catalogs of Die Deutsche Bibliothek to determine literary warrant, but the team preparing the Vietnamese translation consults the catalog of the National Library of Vietnam. Accordingly, these two groups will find different literary warrants for topics of local interest. Dewey copes with these differences in literary warrant through expansion and abridgment.

3. Checking Literary Warrant and Modifying the Classification to Match

The primary source for checking literary warrant was BUBL⁴. Two other sources were used for cross-checking: Canadian Information By Subject⁵ and KidsClick!⁶. BUBL is an Internet-based information service for the UK higher education community and is maintained and operated by the Centre for Digital Library Research of the University of Strathclyde, Glasgow, Scotland. BUBL provides access to approximately 12,000 Internet resources. Canadian Information By Subject is an information service developed by the National Library of Canada. The service provides about 10,000 links to Internet resources about Canada. KidsClick! is a web guide for children created by librarians at the Ramapo Catskill Library System, Middletown, New York. The KidsClick! database provides access to about 6,400 Internet resources. All of these sites use Dewey to provide subject access to resources.

The classification numbers in Abridged Edition 14 were used as a starting point. The abridged edition of DDC is designed to provide the level of detail needed to classify the materials in general collections of 20,000 titles or fewer. The initial sample was limited to 500-559 (Natural sciences and mathematics, not including life sciences). The Abridged Edition class was truncated where literary warrant was lacking. We observed the rules for maintaining true abridgment: the truncated number for a topic is always the same as the full number for the topic, except shorter. For example, the planet Uranus does not need its own number; hence the number for that topic has been truncated to 523.4 Planets from the Abridged Edition number for Uranus, 523.47. For this project there is no need to display all the three-digit numbers of the standard Dewey summary. Thus where literary warrant was lacking, we cut back to one or two significant digits, e.g., cutting 521 Celestial mechanics back to 520 Astronomy and allied sciences. We call

this true abridgment because logically the number 520 is 52, the same as 521 except shorter.

Usually the cross-checking in Canadian Information By Subject and KidsClick! did not affect the outcome. The two exceptions were numbers that had literary warrant in KidsClick! but not BUBL: 507.8, a number much used for science fair projects, and 513, the number for arithmetic.

Where BUBL included a category normally expressed by a number built with a standard subdivision that did not appear in Abridged Edition 14, we added the built number. Both the numbers built with standard subdivisions that we added and those that appear in the schedules would have minimal terminology for matching against electronic resources if we did not make a special effort to enrich the terminology: the editorial rules for the Dewey Relative Index prohibit extensive indexing of numbers built with standard subdivisions. To enrich the terminology in the records for numbers built with standard subdivisions (both those we added and those included in Abridged Edition 14 that had literary warrant in our sample), we added Relative Index terms taken from the relevant Table 1 Standard Subdivisions records. We combined those terms with Relative Index terms from the base number for concepts that approximate the whole of the base number. The terms selected appear as the caption or in class-here notes, or are synonyms of those terms. For example, in adding terminology for the built number 526.06, we used the following Relative Index terms from 526 Mathematical geography:

- Cartography
- Map drawing
- Map making
- Mathematical geography

To these terms we added the following subentries:

- administrative reports
- associations
- intergovernmental organizations
- international organizations
- membership lists
- nongovernmental organizations (international agencies)
- organisations
- organizations
- societies

Thus the record for 526.06 has Relative Index entries like the following:

- Cartography--administrative reports
- Cartography--associations
- Cartography--intergovernmental organizations

Most of these subentries were taken from the Relative Index entries for notation 06 from Table 1. A few terms were added from references in the Relative Index (e.g., "societies") and a few were added for British spelling (e.g., "organisations"). Only the organizations part of T1--06 Organizations and management was given because the literary warrant in BUBL was for organizations, not management. A work on management applied to mathematical geography would stay in the base number 526.

We did not use the term "Geodesy" from 526 in the record for 526.06, for example, because that topic does not approximate the whole of 526. It appears in an including note at 526 in the Abridged Edition. The notes for 526 are shown in Table B. In DDC 22, geodesy is classed in 526.1, a number not in the Abridged Edition; in DDC 22, standard subdivisions for geodesy are added to 526.1, not to 526, e.g., organizations devoted to geodesy 526.106. If we leave works on organizations devoted to geodesy in 526, we maintain true abridgment; the numbers 526 and 526.106 are the same except in length. Our sample contains no works focusing only on geodesy; hence this not a problem.

If there were a substantial number of works on geodesy and on organizations devoted to geodesy in our sample, we would expand beyond Abridged Edition 14 provisions and supply the DDC 22 numbers 526.1 and 526.106. In the sample that we have done so far, we have found no need to expand beyond the provisions of Abridged Edition 14. Table A below shows the classes 526-529 before and after checking literary warrant. The class numbers as represented in Abridged Edition 14 are shown in Column 1. Class numbers that were truncated after checking literary warrant are shown in regular type in Column 2. Built numbers and class numbers that remained unchanged are shown in bold type.

A14	A14 LW	Caption
526	526	Mathematical geography
526.022	526	Illustrations, models, miniatures
	526.06	built number added
	526.06	Built number for organizations devoted to mathematical geography
526.3	526	Geodetic surveying
526.9	526.9	Surveying
526.9092	526.9	Surveyors
527	52	Celestial navigation
528	52	Ephemerides
529	529	Chronology

Table A. Classes 526-529 before and after literary warrant checking.

4. Creating a Version of DDC for Automated Classification

The machine service that performs automated classification consists of two major components: a database of concepts used to classify a document, and software that

generates a prioritized list of concepts that roughly characterize the content of the document. Records in the database are constructed from the following elements of a class:

- Class number
- Caption
- Superordinate hierarchy
- Notes that describe what is found in a class⁷
- Relative Index entries
- Mapped terminology

For truncated numbers, information from the truncated class was merged into the record for the longer number. For example, the caption, notes and Relative Index entries for 526.3 were added to the record for 526 as shown in Table B. Mapped terminology is omitted in the example. For built numbers that did not appear in Abridged Edition 14, new records were created as described above. To test the effectiveness of the database, a sample of resources, in the range 500-559, from each of the three web sites studied will be automatically classified by the service. The resulting class numbers will be manually evaluated for accuracy.

5. Summary

This paper presented the results of an exploratory study to determine literary warrant for topics in 500-559 (Natural sciences and mathematics, not including life sciences) in electronic resources. By looking at web sites for literary warrant and modifying the classification to match, a version of the DDC was created that accommodates the topics found on three web sites of 6,000 to 12,000 resources. The modified version of the DDC is being tested as the source database for machine classification services. If the results of the automated classifications are satisfactory, the approach may be extended to other web sites that vary in size and scope.

	Data elements from 526	Data elements from 526.3
Class number	526	
Caption	Mathematical geography	Geodetic surveying
Hierarchy	Science Astronomy	
Notes	Including geodesy, gravity determination, latitude, longitude; Class here cartography, map drawing, map making	Surveying in which curvature of the earth is considered; Including leveling, triangulation
RI entries	Cartography Equator Geodesy Gravity determination--geodesy Latitude	Geodetic surveying Leveling (Surveying) Triangulation

	Data elements from 526	Data elements from 526.3
	Longitude Map drawing Map making Mathematical geography Tropic of Cancer Tropic of Capricorn	

Table B. Database record for 526 with data elements from 526.3

Notes

¹DDC, Dewey, Dewey Decimal Classification, WebDewey, and WorldCat are registered trademarks of OCLC Online Computer Library Center, Inc.

² OCLC's cataloging service.

³ A PDF version of the Glossary reprinted from volume 1 of DDC 22 is available at <http://www.oclc.org/dewey/versions/ddc22print/>

⁴ The BUBL service is accessible at <http://bubl.ac.uk/>

⁵ Canadian Information by Subject is accessible at <http://www.nlc-bnc.ca/caninfo/esub.htm>

⁶ KidsClick! is accessible at <http://sunsite.berkeley.edu/KidsClick!/>

⁷ These include definition and scope notes, former-heading notes, variant-name and former-name notes, class-here notes, and including notes (Chan and Mitchell, 2003, 22-25).

References

- Chan, L. and J. S. Mitchell. 2003. *Dewey Decimal Classification: Principles and Application*. Dublin, Ohio: OCLC.
- Beall, J. 2003. Approaches to expansions: case studies from the German and Vietnamese translations. Presented at the Classification and Indexing – Workshop, World Library and Information Congress: 69th IFLA General Conference and Council. 1-9 August 2003, Berlin. <http://www.ifla.org/IV/ifla69/papers/123e-Beall.pdf>
- Mitchell, J. S. 2001. Relationships in the Dewey Decimal Classification System. In *Relationships in the Organization of Knowledge*, edited by C.A. Bean and R. Green, 211–226. Dordrecht: Kluwer Academic.
- Mitchell, J. S. 2003. DDC 22 offers many updates to Dewey users worldwide. *OCLC Newsletter*, July No. 261. Also available at: <http://www5.oclc.org/downloads/design/e-newsletter/n261/ddc22.htm>
- Shafer, K., Thompson, R. and V. Tkac. 1997. Scorpion: Dewey Database Design. http://orc.rsch.oclc.org:6109/dewey_db_design.html
- Vizine-Goetz, D. 2001. Dewey in CORC: Classification in Metadata and Pathfinders. *Journal of Internet Cataloging*, (4), 1/2, 67-80. Also published in *CORC: New Tools and Possibilities for Cooperative Electronic Resource Description*, edited by K. Calhoun and J.J. Riemer, 67-80. New York: Haworth

Appendix 1.

PowerPoint presentation available at:

<http://www.oclc.org/research/presentations/vizine-goetz/isko2004.ppt>