Some Temporal Aspects of Indexing and Classification: Toward a Metrics for Measuring Scheme Change

Joseph T. Tennis
University of Washington
Information School
Seattle, WA
98195 USA
001 206 616 2542
jtennis@uw.edu

Katherine Thornton
University of Washington
Information School
Seattle, WA
98195 USA
001 206 616 2542
thornt@uw.edu

Andrew Filer
Datamarcx
Seattle, WA
98102 USA
001 612 709 5417
afiler@gmail.com

ABSTRACT
In this paper we discuss the temporal aspects of indexing and classification in information systems. Basing this discussion off of the three sources of research of scheme change: of indexing: (1) analytical research on the *types* of scheme change and (2) empirical data on scheme change in systems and (3) evidence of cataloguer decision-making in the context of scheme change. From this general discussion we propose two constructs along which we might craft metrics to measure scheme change: collocative integrity and semantic gravity. The paper closes with a discussion of these constructs.

Categories and Subject Descriptors
H.3.1 **[Information Storage and Retrieval]**: Content Analysis and Indexing.

General Terms
Management, Measurement, Documentation, Performance, Design, Human Factors, Languages, and Theory.

Keywords
Indexing, classification, time, scheme change, constructs for measurement.

iConference 2012, February 7-10, 2012, Toronto, ON, Canada.
ACM 978-1-4503-0782-6/12/02.

1. INTRODUCTION
In this paper we discuss the temporal aspects of indexing and classification in information systems. Basing this discussion off of the three sources of research of scheme change: of indexing: (1) analytical research on the *types* of scheme change and (2) empirical data on scheme change in systems and (3) evidence of cataloguer decision-making in the context of scheme change. From this general discussion we propose two constructs along which we might craft metrics to measure scheme change: collocative integrity and semantic gravity. The paper closes with a discussion of these constructs.

2. RATIONALE
If we define information systems as those that store, organize, and make accessible documentary material, then the contemporary information system landscape is one made up of hybrid systems. Some are hybrids of digital and non-digital information (legacy systems) others are made up of different data types and different metadata schemas. These, like long-standing social science surveys [1], are the subject of applied and basic research projects. That is, there is a pressing need to understanding how to manage such hybridity. A key component to solving the hybridity problem is to design for time and temporal effects in indexing languages and classification systems. If we are able to design accounting for the fact that the system, its organization, semantics, and encoding will change, we can maintain the design mandates of systems. It is possible that such maintenance would not only ensure the viability and usefulness of the information system in a consistent manner for users, but might also save money, by not requiring major initiatives to fund retrospective conversion. In short, we want to maintain system viability over time.

How, then, do we account for the temporal aspects of indexing languages and classification systems? We must describe the current states of the languages and systems, chart how they change, and then propose design requirements to link the new state to the old state, thereby maintaining the functionality proposed by the original system.

3. SCHEME CHANGE
3.1 Structural, Word-Use, and Textual Change

When we examine the way indexing languages and classification schemes (schemes) change over time we can see three major types of change, each with subtypes [7].

Scheme change occurs in three general categories: structural change, word-use change, and textual change. Structural change deals with the relationship structures in schemes and how editors alter them. Word-use change affects definitions, word forms, lead-in terms, etc. Though both structural changes and word-use changes are semantic, the latter do not explicitly affect relationship structures. Textual changes can affect both structural and word-use changes. Textual changes are changes in the interpretation and assignment of classes to types of documentary material. The first two fall into the purview of the editors, while the third falls to both the editor and indexers. The following characteristics of scheme changes are adapted from [3-6].

3.1.1 Structural Change

Structural changes affect a user's navigation through the scheme. Structural changes affect the semantics of a scheme because they change the relationships that obtain between classes in that scheme. Structural change falls into five basic changes.1 The five basic changes are:

• Addition of a new class (deletion of class)

• Change in synonym structure (use eugenics to lead to both genetics and psychology)

• Change in equivalence structures (e.g., USE and/or USED FOR)

• Assignment of class to another group in the hierarchy (medicine becomes and applied science rather than a natural science)

• Addition or elimination of associative relationship (e.g., RT).

The degree to which these changes affect indexing or classification is dependent on the purpose and structure of the scheme before the change. That is, if a scheme is a thesaurus built on principles of mutual exclusivity (only one place for each concept–no overlap) then these changes are dramatic. If the scheme is not built on principles of mutual exclusivity, then navigation is hindered, but not confounded through these changes. In either case, it is desirable in a digital environment to track these characteristics of change in order to mange the meaning communicated through indexing and retrieval process.

3.1.2 Word-Use Change

The second type of change is word-use change. Word-use changes do not affect navigation through the structure. They are changes that preserve the structure of a scheme, while adding or replacing words. This may affect indexing and classification practice, but it does not affect the scheme structurally. Word-use changes are:

• New word used as lead-in (adding blacks to the lead-in list for African Americans)

• New synonyms added (replaced one for one, for example, genetics for eugenics)

• New preferred class added (preferring homosexuals over inverts)

• Change in definition of class (changing the definition of computer from "a calculating machine" to "electronic device for storing and processing data, typically in binary form")

Like structural changes, combinations of these changes can occur. The effect of word-use on scheme versioning is powerful. In our example of eugenics, the lead-in terms, the synonyms, and the definition all affect the use of the class. *Eugenics* has been a concept that has affected a number of areas of science, social science, and philosophy. How words are used to present this concept affect the way it will be used by the indexer. One can also imagine a scenario where a class may be present in the scheme, but not used because of the definition. If this remains a constant in the use of the scheme, then this has ramifications for the structure. The class may disappear for example if its not used. So it is not structure alone that affects structural changes. Textual changes also affect structural changes, as well as word-use changes.

3.1.3 Textual Changes

Textual changes are changes in the relationships between texts and a version of the scheme. There are two primary types of textual changes.

The first is textual warrant change and the second is the document-set change. Textual warrant is a term that is close to literary warrant–but does not mean the same thing. Textual warrant is the combination of all texts (literature of the field, user studies, search logs, checklists, etc.) that would be used to create a value or relationship in a scheme. Soergel calls these sources and authorities [5]. Any change in this collection of texts results in a change of the evidence considered when managing the scheme, and hence managing changes to the scheme.

The second kind of textual change is the document-set change. In this case a set of documents has been indexed and given a value (for example, 575.6). This set will change as the scheme changes, and therefore shifting the representation power of the scheme. So the texts once classed under 575.6 are not the same kind of texts, because the relationship between the document set and the value has changed (In bringing up this concept of a document-set and its shifting representation, I am invoking an analytical device similar to Melanie Feinberg's [2], though not identical in use; they are similar in composition).

4. CATALOGUERS' DECISIONS OVER TIME

Given all these types of change, in the abstract, we want to know how change effects decisions made during subject cataloguing. We know that cataloguing is a negotiated process. We also know that given a choice between user term, author terms, and terms in indexing languages, priority is often given to the indexing languages used by librarians [8]. What then, does cataloguing practice look like when a subject changes? In order to answer this question we collected two types of data. We followed one subject, *Eugenics*, through all extant editions of the Dewey Decimal Classification (DDC), from its first appearance in 1911 to the 22nd edition that was published in 2003. We could then see what classes were available to the cataloguer. The second set of data were gathered using Z39.50 protocol, harvesting MARC records from 572 catalogues that both (1) used *Eugenics* as a first subject heading (in the 650 field of the MARC record, the subject added entry for topics) [10], and (2) used the DDC in the 082 field of the MARC record. After removing duplicate records we were left with 477 records. These records are evidence of cataloguer decision-making. We are now able to compare the data gathered from the editions and the data gathered from the Z39.50 protocol, and begin to make some statements about how scheme change effects cataloguing.

Eugenics is generally understood to be the applied science or the biosocial movement which advocates the use of practices aimed at improving the genetic composition of a population. Usually refers to human populations [9].

In the next sections we present both sets of data, then a combined visualization of the data, and we close with a proposition for semantic gravity, which we believe will move us toward a quantifiable effect of scheme change.

5. EUGENICS IN DDC FROM 1911-2003

What follows is a table, drawn from all unabridged versions of DDC in the English language that referenced *Eugenics* in the Relativ Index – the terminological access to concepts and classes in the DDC. The Relativ Index is useful to cataloguers because it guides them to make decisions about the context of the class, whether it is a science or a social science for instance. This is because the DDC is organized first by discipline, then by topic within disciplines. The table below shows where it was possible to place the topic *Eugenics* in the DDC from the 7th edition published in 1911 to the 22nd published in 2003.

Table 1a. *Eugenics* in the DDC 1910s-1950s

Class Numbers For *Eugenics* in the Relativ Index	575.1 575.6 613.94	575.1, 613.94	136.3 159.922 3 575.1, 613.94	136.3 301.323 364.301 8 364.42, 575.1 613.94	364.42 613.94
Edition Numbers	7th, 8th, 9th, 10th	11th, 12th,	13th	14th	15th 16th
Decade	1910	1920	1930	1940	1950

Table 1b. *Eugenics* in the DDC 1960s-2010s

Class Numbers For *Eugenics* in the Relativ Index	613.94	174.25 363.92 363.98 364.4 364.42, 613.94	174.25, 323.97, 362.36, 363.92, 363.97, 363.98, 364.42, 613.94	174.25, 362.36, 363.92, 363.97, 364.4, 613.94	176 323, 362.36, 363.92, 363.97, 364.4, 613.94
Edition Numbers	17th	18th, 19th,	20th	21st	22nd
Decade	1960	1970	1980	1990	2000

As can be seen above, there was no entry in the Relativ Index of the 17th edition of the DDC for *Eugenics*, though it does appear in the class schedule as *Eugenic Practices* so it is retained here. However, in all other editions cataloguers would find the term *Eugenics* in the Relativ Index.

We have placed the numbers close to one another to approximate the spatial relationships of the topic as imposed by the ordinal notation of the classification scheme. We can observe two structural changes: addition and deletion. We would have to probe deeper to see others.

This table represents what is possible to say with regard to *Eugenics* in the DDC. We can now look at the results of cataloguers' decision and compare these two.

6. EUGENICS IN THE HANDS OF CATALOGUERS

Drawing on data gathered from the Z39.50 harvest of MARC records that had *Eugenics* in the first 650 field and had a DDC number in the 082 field, we are able to line up books catalogued on eugenics by estimated date catalogued (using the LCCN in the MARC record), date published, DDC number. We also have counts of duplicates. That is we know how many books are in our collection that have the same title, publication date, and class number.

A small subset of the data follows. Here we see books classed using DDC before the Relativ index had the term *Eugenics.*

Table 2. Books Catalogued before 1911

Year Catalogued	Publication Date	DDC	Total Count
1902	1902	901	1
1907	1907	173.5	2
1908	1839	613.94	2
1909	1875	176	1
1909	1908	573	1
1909	1909	613.9	2
1909	1909	575.1	1
1910	1883	155	4
1910	1910	612.6	1

We can see that the books are classed in history, philosophy, psychology, medicine, and science. The number 575.1 is noteworthy here because it is linked to the topic *Genetics* and is used for *Eugenics* in the 1911 edition of the DDC. The status of *Eugenics* as a science offers a challenge to cataloguer and those making decisions about the classes in the DDC.

Table 3. Books Catalogued in 2010

Year Catalogued	Publication Date	DDC	Total Count
2010	2010	363.9209409041	2
2010	2009	370.15	1
2010	2010	509.470904	1
2010	2010	176	1

Above we see books catalogued in 2010 in our data. We see some changes and some stasis. There is still science and philosophy, but we add social science to disciplines that contextualize the topic. All of these are from the 22nd edition of the DDC. So though these all have *Eugenics* as the first subject

heading in the MARC 650 field we can see that one is considered a science (509.470904) not a social science science.

7. CLASSES POSSIBLE AND CLASSES USED

When we line up the two data sets we can produce graphs that allow us to see where in the scheme (1) classes are available for the topic *Eugenics* and where cataloguers placed books on the topic *Eugenics*. Again, books are considered "on the topic *Eugenics*" if *Eugenics* appears in the first 650 field of the MARC record.

What follows are five charts, four of which show a century (e.g., 500-600) of the DDC classes. We note our observations.

Below is a chart for the 500s in the DDC. We are charting here, where a book as classed and when, and how it compares to the DDC schedules for *Eugenics.* What we expect to see is that cataloguers follow the current edition of the scheme and only class books where it is possible to class books. That is, we should see small diamonds appear after large squares, and they should not appear after large Xs.

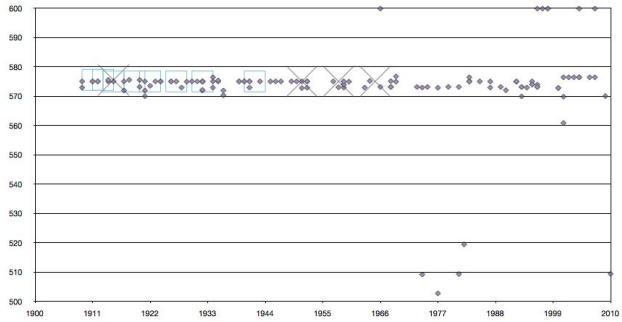


Figure 1. The 500s – Science - in DDC. The small diamonds are books classed, the empty squares are possible classes, and the large Xs are discontinued classes. © Joseph T. Tennis

We see a square for each time the DDC offers a class for the topic *Eugenics.* We can see that they offered classes for *Eugenics* from 1911 until the late 1942. We can see in this chart that though classes are formally discontinued (marked by the X) in the DDC they are still used by librarians. This means that cataloguers are deciding to respect the collocative power of their collection over and above the collocative power of the new edition of a classification scheme.

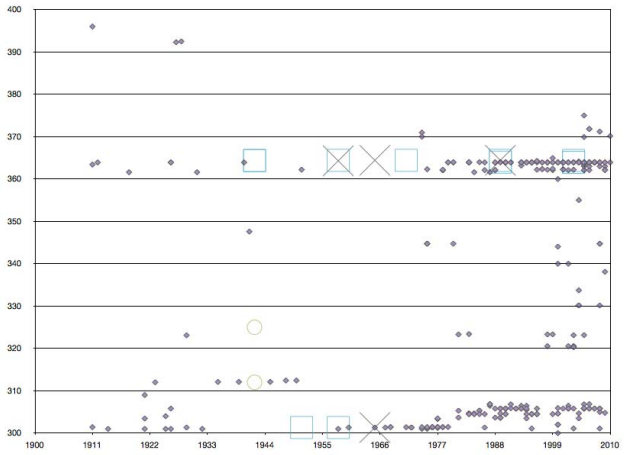


Figure 2. The 300s – Social Science - in DDC. The small diamonds are books classed, the empty squares are possible classes, the circles are see also references, and the large Xs are discontinued classes. © Joseph T. Tennis

We see here the same data, but for the Social Sciences. There is more scatter here, but we observe the same phenomenon, that cataloguers continue to use classes even when they are not provided in the classification scheme.

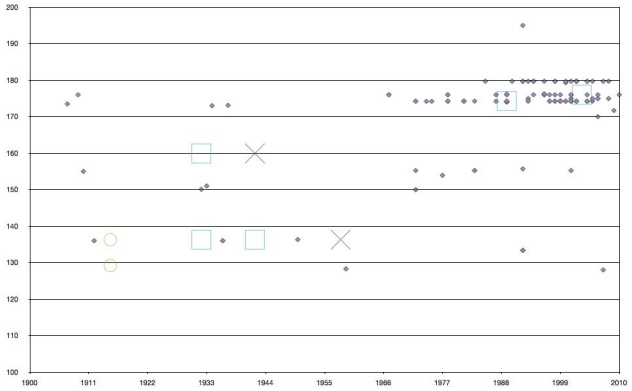


Figure 3. The 100s – Philosophy – in DDC. The small diamonds are books classed, the empty squares are possible classes, the circles are see also references, and the large Xs are discontinued classes. © Joseph T. Tennis

In Philosophy we see a pattern closer to our prediction. In this case we see that a class surfaces, see the upper right-hand side, and then instances of books classed follows. We also see discontinued classes being honored by an absence of books classed. Useful Arts / Technology and Applied Science is next.

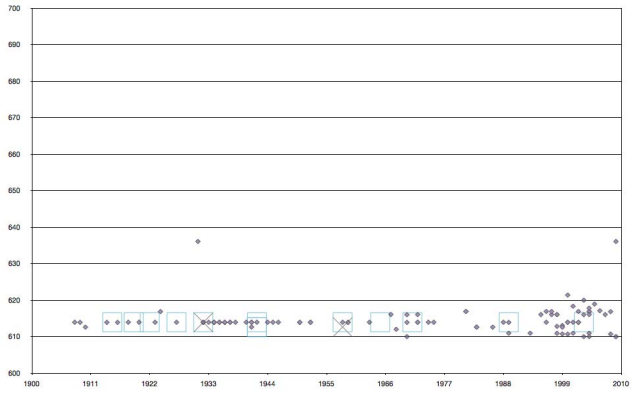


Figure 4. The 600s – Useful Arts / Technology and Applied Science – in DDC. The small diamonds are books classed, the empty squares are possible classes, the circles are see also references, and the large Xs are discontinued classes. © Joseph T. Tennis

With Useful Arts / Technology and Applied Science we see more coherence and less wide-ranging interpretation until the end of the Twentieth Century. This may be a direct reflection on the literary warrant. That is, we may see quite a diverse set of interpretations with regard to *Eugenics* as a topic in the discipline. Cataloguers may recognize the topic, but not be sure where in the disciplines the book belongs.

The following table is a composite of the previous four. It shows the increase in publication, and the consistent decision by cataloguers to class books in deprecated classes. It is also appended at the end of the paper (Appendix A.).

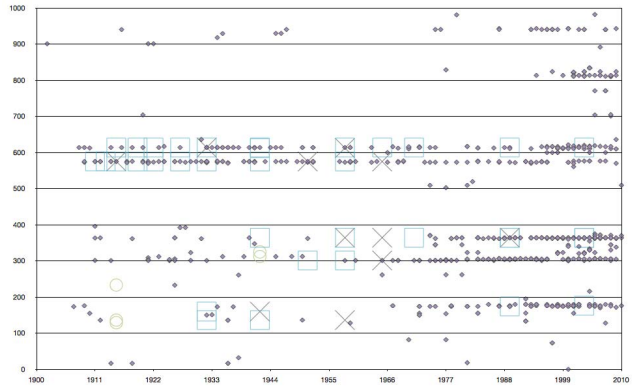


Figure 5. Composite of all classes in DDC. The small diamonds are books classed, the empty squares are possible classes, the circles are see also references, and the large Xs are discontinued classes.

From this data we do not observe decision-making that comports with the semantics of the current edition of the scheme. So, though the DDC is released, cataloguers do not always follow the most current version in their classifying, at least there is a discrepancy between following a prescribed scheme for a while, and then not following the prescribed scheme at a later date. And while it is possible to track which edition of DDC they are using if they provide that data in the MARC record, it is commonly only available for editions 19 forward. So though we could look only at the subset of those data (which would be 1979-present), we would not have a complete picture of the life of this topic in the scheme.

8. MOVING TOWARD METRICS

From this exploratory data analysis we can begin to characterize this phenomenon by establishing constructs that frame a quantifiable aspect of this phenomenon. However, we need to do this with the full context. So we need comparative data from other classes and other timelines of cataloguer decision-making. We are moving toward that, and we are in the process of analyzing other topics.

In the interim we propose two theoretical constructs to help guide and give meaning to the characterization of scheme change over time.

First, since we are dealing with systems for information browsing and retrieval we are concerned with collocation, that is, pulling together kinds of literature. In this case we are trying to pull together books on *Eugenics* into the disciplines of Science, Social Science, Philosophy, or Useful Arts / Technology and Applied Science. If we are to maintain the functional requirements of systems like this over time, then we have to measure the *collocative integrity* of a scheme as it changes. For example, is there a point at which we can say, based on empirical data and a threshold measure that the class, say, 155, is not useful to collocate subject? Is there a measure we can use to say this class has lost its collocative integrity?

Second, we can see from the charts above that cataloguers often favor a class that is no longer provided by the updated classification scheme. This may be because there are books already in the collection that are *about* the topic, in this case *Eugenics.* We know that cataloguers will privilege their perception of shelf collocation over other semantic factors when subject cataloguing [8]. This raises the question whether or not there is a *semantic gravity* that a body of classed books has that pulls a soon-to-be classed book toward that class number, and away from the revised classes possible in the newly updated scheme. Does this semantic gravity have a numeric function, a metric?

In both of these theoretical constructs we have to weigh the counter arguments. Who cares about the collocative integrity of a classification scheme? What if there is no consistent phenomenon that resembles a semantic gravity? In both of these cases we have to measure our skepticism against the raison d'être of the classification scheme: to pull together kinds of documents for a user, independent of vocabulary and in a single place. That is, where terminological search matches words to concepts, classification is supposed to display the set of those objects relevant to the search.

If this is a design requirement that is to be maintained over time we must consider the both how the classification scheme collocates and what might deter the classification scheme from collocating. So we must consider this question in some fashion. And it appears that we have a good start in collecting data to support an empirical approach to understanding these problems that creep into classification schemes over time.

9. ACKNOWLEDGMENTS

My thanks to Katherine Thornton, Andrew Filer, Gary Gao, Tod Robbins, and Monica Caraway for their contributions to the

empirical data collection, organization of research materials, visualizations, and general comments. The errors in this paper are solely those of the author.

10. REFERENCES

[1] Metadata for Long-standing Large-Scale Social Science Surveys (META-SSS) - NSF 11-583. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504705

[2] Feinberg, M. (2005) Expression of feminism in three classifications. In Andersen, Jack, ed. *Proceedings 16th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research,* October 29, 2005, Charlotte, NC.

[3] Beghtol, C. (1986). Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. Journal of Documentation 42(2): 84-113.

[4] Ranganathan, S. R. (1967). Prolegomena to library classification. 3rd edition. Bombay: Asia Publishing.

[5] Soergel, D. (1974). Indexing languages and thesauri: construction and maintenance. Los Angeles: Melvile.

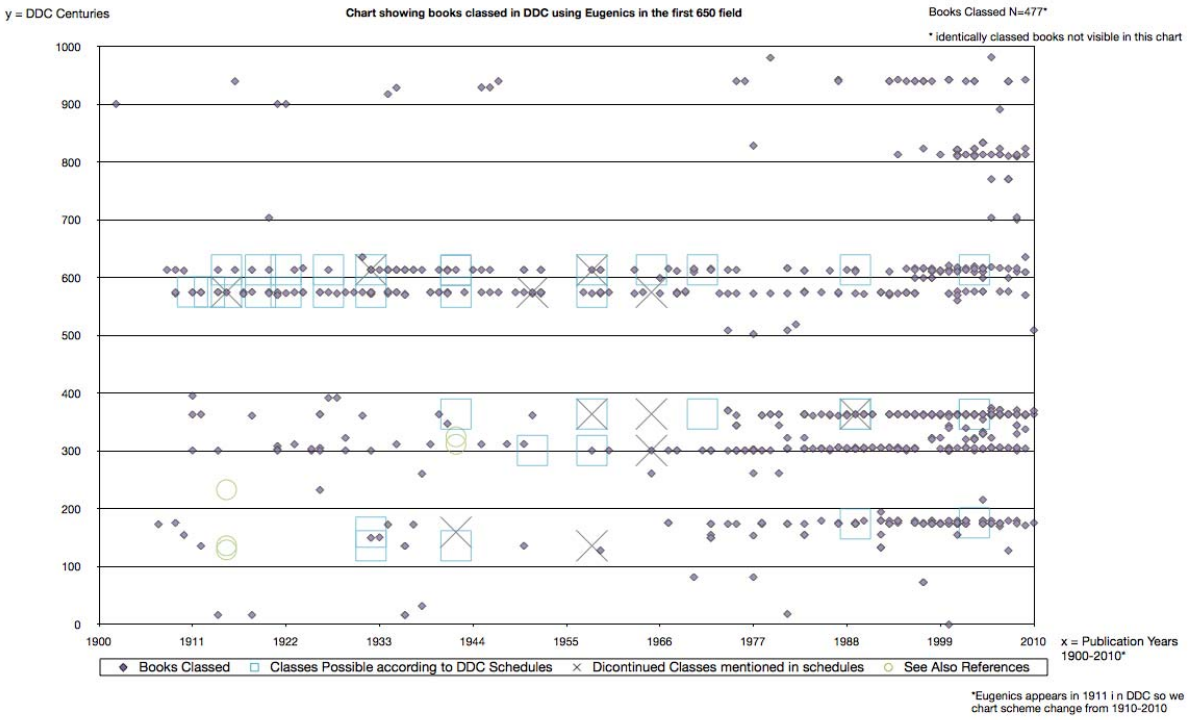
[6] Aitchison, J. Gilchrist, A, and Bawden, D. (2000). Thesaurus construction and use: a practical manual. London: Aslib.

[7] Tennis, J. T. (2007). Scheme Versioning in the Semantic Web. In *Cataloging and Classification Quarterly*. 43(4/3): 85-104.

[8] Šauperl, A. (2003). Cataloger's Common Ground and Shared Knowledge. Journal of the American Society for Information Science and Technology 55(1): 55-63.

[9] Unified Medical Language System (Psychological Index Terms) at the National Library of Medicine. http://ghr.nlm.nih.gov/glossary=eugenics

[10] Library of Congress. (2007). 650-Subject Added Entry – Topical Term. http://www.loc.gov/marc/bibliographic/bd650.html



Appendix A. Larger Composite of Eugenics Classes and Classed Books in DDC