Tennis, J. (2013). Collocative Integrity and Our Many Varied Subjects: What the Metric of Alignment between Classification Scheme and Indexer Tells Us About Langridge's Theory of Indexing. NASKO, 4(1). Retrieved from http://journals.lib.washington.edu/index.php/nasko/article/view/14660

Collocative Integrity and Our Many Varied Subjects: What the Metric of Alignment between Classification Scheme and Indexer Tells Us About Langridge's Theory of Indexing

Joseph T. Tennis
University of Washington Information School

As the universe of knowledge and subjects change over time, indexing languages like classification schemes, accommodate that change by restructuring. Restructuring indexing languages affects indexer and cataloguer work. Subjects may split or lump together. They may disappear only to reappear later. And new subjects may emerge that were assumed to be already present, but not clearly articulated (Miksa, 1998). In this context we have the complex relationship between the indexing language, the text being described, and the already described collection (Tennis, 2007). It is possible to imagine indexers placing a document into an outdated class, because it is the one they have already used for their collection. However, doing this erases the semantics in the present indexing language. Given this range of choice in the context of indexing language change, the question arises, what does this look like in practice? How often does this occur? Further, what does this phenomenon tell us about subjects in indexing languages? Does the practice we observe in the reaction to indexing language change provide us evidence of conceptual models of subjects and subject creation? If it is incomplete, but gets us close, what evidence do we still require?

To address these questions we documented how different subjects changed over time in the Dewey Decimal Classification (DDC). For example, we marked where one could class the topic EUGENICS. In 1911 it is a biological science. However, it can no longer be classed in 575.6, which is now the class for REPRODUCTIVE PARTS OF PLANTS. We collected this data from 1876-2003.

Then, using the Z39.50 protocol we downloaded bibliographic records from 665 libraries using the DDC. The libraries were chosen from a list of those offering Z39.50 access to their catalogues and they also used the DDC. The libraries are in North America, South Africa, Taiwan, and Europe. We arranged the records according to Library of Congress Control Number. This number has a date built into it. See examples below with date highlighted in bold. Before the year 2000 LCCN dates were two digits; after 2000 they were four digit years.

LCCN	Date Inferred
68098003	1968
2004049123	2004

Table 1. LCCN Dates

The date signifies when the bibliographic description was done at Library of Congress. This gives us an approximate time of cataloguing. We chose records that have our subject term in the MARC 650 field. So in our example above we arranged bibliographic records that had EUGENICS in the first MARC 650 field, with the assumption that we would see cataloguers class this where EUGENICS was available in the DDC.

From here we can ask how many cataloguers agreed with DDC and how many did not. We can also ask how many documents were classed in outdated numbers. Furthermore, we can observe trends in agreement and disagreement. These observations can be quantified, and that is what gives us the concept and metrics of *collocative integrity*. The measures of collocative integrity are given in the table and figure below. Here we show you where books on EUGENICS are classed. They are either in a class, outside of a possible class, or in an outdated class number. Over all only around 28% of the books on EUGENICS were classed where the DDC provided explicit semantic matches.

Eugenics

	In	**Out**	**Old**
1899-2003	244	623	14
Percent	~28%	~71%	~1%

Table 2. Counts and Percentages of Eugenics Books Classed In, Out, and in Old DDC Numbers

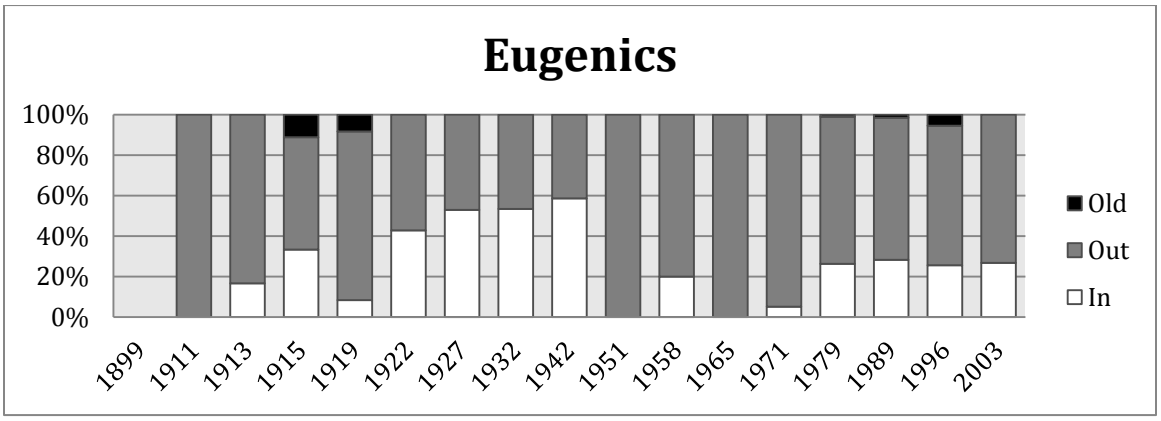

Figure 1. Visualization of the Percentage of Eugenics Books Classed In, Out, and in Old DDC Numbers (NB: 1899 has no data so there is no bar)

The measure of collocative integrity can be used in a number of ways. For example, we can see, at one point in time, where subject headings match well with the subjects in the scheme. We can see when, and how often, old classes are used. We might also be able to help indexers improve their practice. It might be that a subject with low collocative integrity can be flagged for potential reassignment. We might also be able to help designers of indexing languages plan revisions based on an optimized conceptualization of collocative integrity. Perhaps there is a benchmark they want to be above in the alignment of indexer and indexing language.

We might postulate that in an ideal sense, as schemes change, the integrity of the collocation of documents on topics remains intact, that we do not jeopardize collocative integrity when we revise and restructure schemes.

As for the potential research impact of measuring collocative integrity we may be able to explore the conceptions of types of subjects and types of subject emergence. Langridge, in analyzing the DDC talks about *forms of knowledge, topics, specializations, forms of*

writing, *forms of thought*, and *forms of text* (Langridge, 1989). The first three mentioned by Langridge are his conceptualization of how something is studied, the object of study, and the combination of the two. So PHILOSOPHY is a form of knowledge and can be applied to a range of topics. THE MIND or DESCRIPTORS IN THE MLS BIBLIOGRAPHY are topics that can be studied from various disciplinary points of view. Specializations are the intersections of these two, commonly called disciplines, but Langridge does not see the term as useful because it conflates specializations and forms of knowledge, where he sees them as distinct. An example of a specialization is The PHILOSOPHY OF MORALS commonly called ETHICS.

We can see in our data what might be evidence of these distinctions. For example, when we chart the collocative integrity of ANATOMY we see a high level of integrity over time. This is because the description of ANATOMY is solidly a MEDICAL SCIENCE or an ART in the DDC. This is because we are commonly writing about the practice of anatomy or the drawing of anatomy. The indexers and the scheme agree based on the structure of this *specialization*. Table 3 and Figure 2 illustrate this.

Anatomy

	In	**Out**	**Old**
1899-2003	1219	666	195
Percent	~59%	~32%	~9%

Table 3. Counts and Percentages of Anatomy Books Classed In, Out, and in Old DDC Numbers

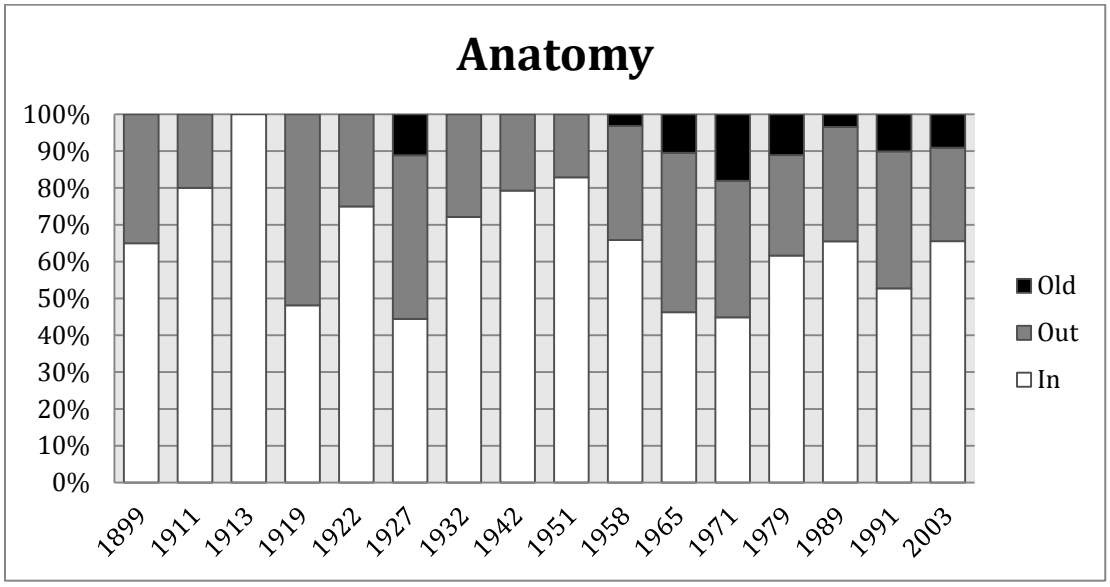

Figure 2. Visualization of the Percentage of Anatomy Books Classed In, Out, and in Old DDC Numbers

We have close to 60% collocative integrity for ANATOMY compared with the 28% of EUGENICS. As a *topic* it is, and can be, studied from a number of *forms of knowledge*, so there is no consistent *specialization* over time. We can show this scatter visually by showing a timeline of classes possible in DDC and showing where cataloguers placed books in or out of those classes. See Figure 3 below.

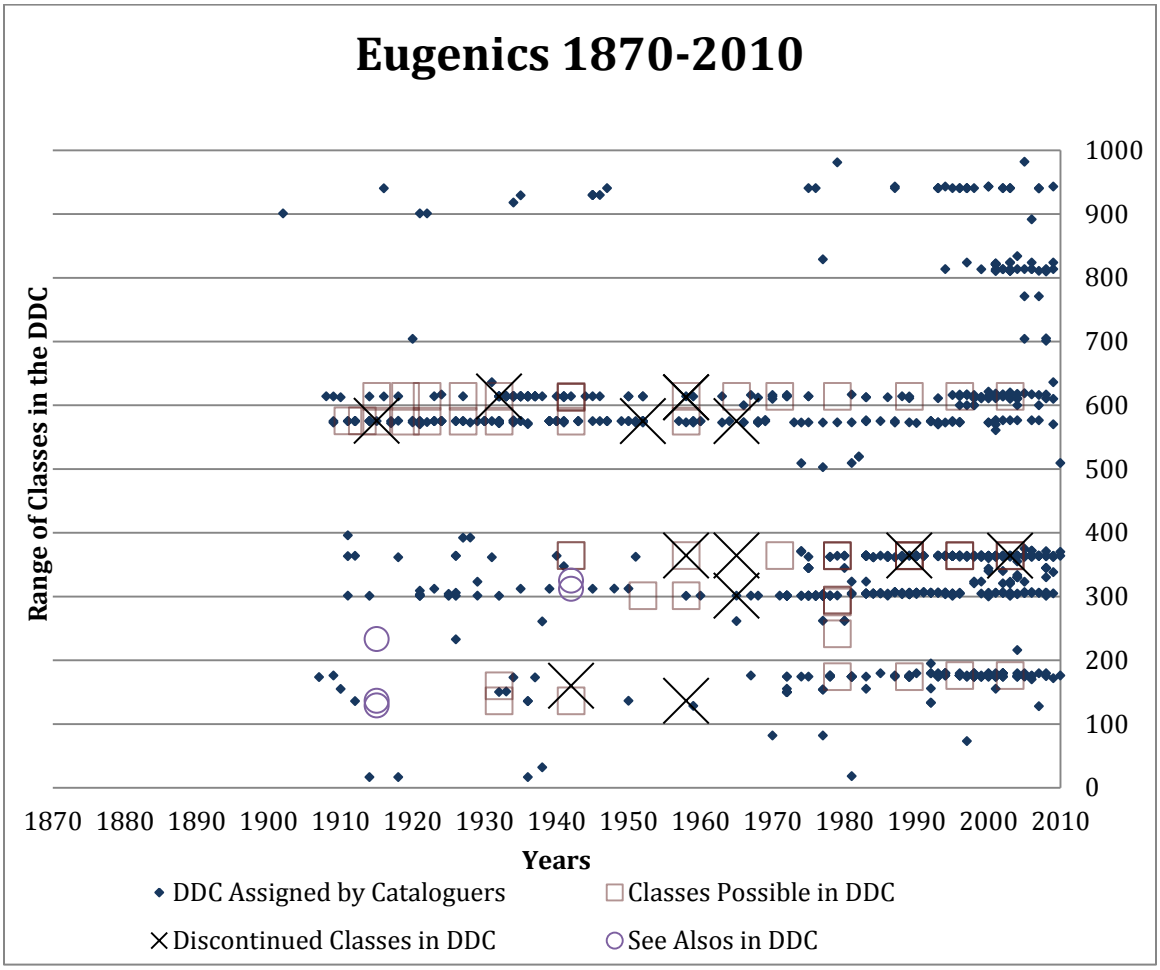

Figure 3. Eugenics classes and books arranged in chronological order and in DDC class number order

Above we see squares that indicate the classes possible in DDC for EUGENICS. The large x's show where the editors of DDC explicitly removed a class from the schedules. The circles are see-also references, and the small diamonds are unique cataloguer decisions. There are 891 unique decisions presented here. Unique decisions for our purposes were records that had the same subject heading, different title, different publication year (if title was the same), and different class number.

We can see a range of forms of knowledge represented both in the classes possible in DDC and in cataloguer decisions. The 100s are philosophy and psychology, 200s are religion, 300s are social sciences, 500s life sciences, 600s applied sciences and useful arts, 700s fine arts, 800s are literature, and 900s are history and geography.

We also see, perhaps, a range of topics and specializations in, for example the 300s and 500s.

We can compare this graph to the one derived from decisions made about books on ANATOMY and GYPSIES.

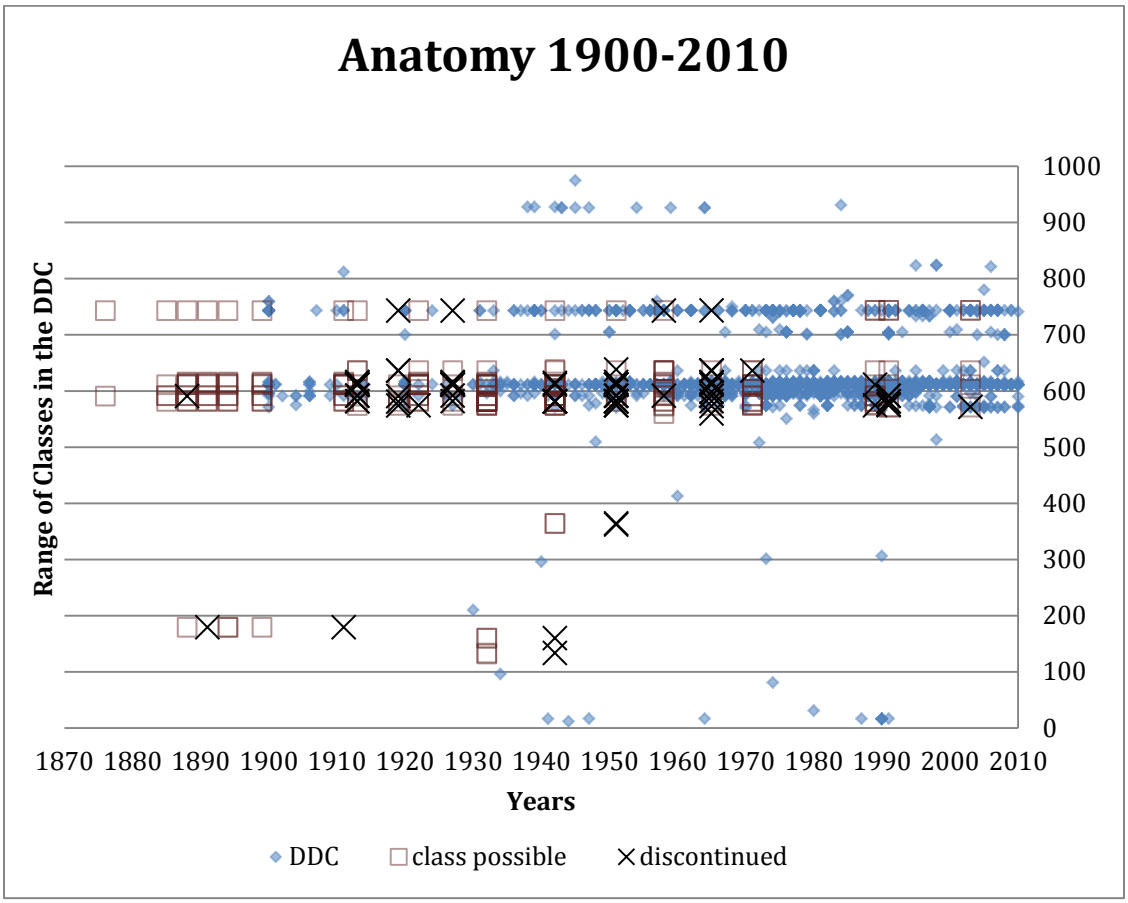

Figure 4. Anatomy classes and books arranged in chronological order and in DDC class number order

The above figure reinforces, in visual form, the measure of collocative integrity present in the cataloguing practice of anatomy books. Cataloguers consistently agree with the DDC and place books on ANATOMY in either applied sciences (600s) or fine arts (700s). Even when DDC introduces natural sciences classes with few exceptions cataloguers agree with the editors of the classification scheme. The other notable difference here is with the scheme and its lack of see-also references for this particular subject.

When we consider GYPSIES as a subject, or rather, as Langridge would describe it as a topic, we see a different set of considerations surface. It looks like both ANATOMY and EUGENICS. As a topic GYPSIES can be considered from different forms of knowledge. This makes it similar to EUGENICS. We see this over time as displayed in Figure 5. However, we also see a disagreement between cataloguers and the prescriptions of DDC. They do not agree as to where GYPSIES belong in the range of classes. This is in part due to the way people are handled in DDC as topics. From 1965 (17th Edition) onward we

get both classes in the schedules for people (and their languages, for instance), but then we also see the editors of DDC move area, ethnicity, race, and nationality, as well as language, to the tables for synthesis to forms of knowledge. There is a similarity between GYPSIES and ANATOMY in the stability of forms of knowledge over time. With the only change coming with the way DDC treats people in the 1960s onward. And we do see some collections privileging older conceptions of GYPSIES as a cultural group with distinct art forms, language, and social customs. We also see at least one *echo class* that resurfaces in 2003 in response to cataloguer work in the 900s.

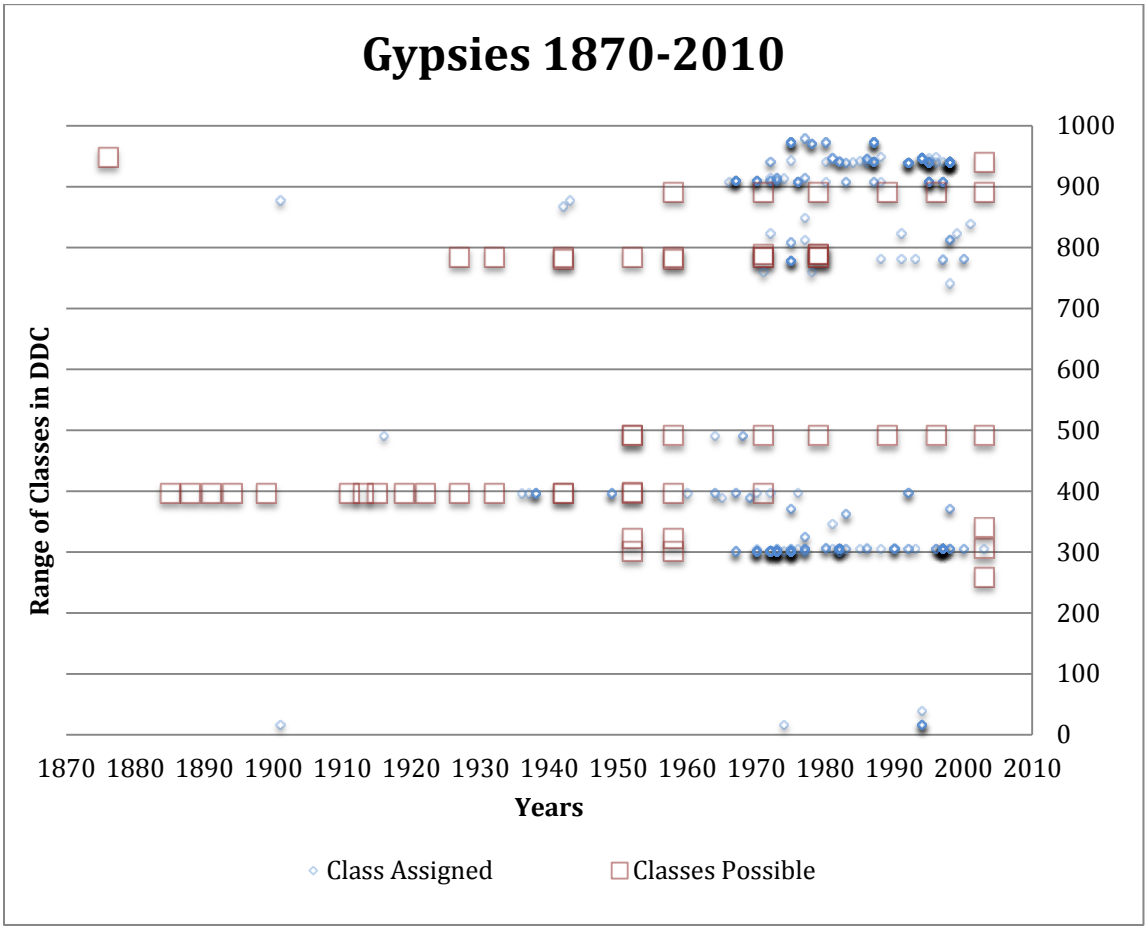

Figure 5. Gypsies classes and books arranged in chronological order and in DDC class number order

The collocative integrity measure for GYPSIES is problematic because from 1965 onward much of the semantics of GYPSIES is derived from number synthesis based on Tables. Our first attempt at calculating this measure can be seen in the table and figure below.

Gypsies			
	In	**Out**	**Old**
1899-2003	17	339	52
Percent	~4%	~83%	~13%

Table 4. Counts and Percentages of Gypsies Books Classed In, Out, and in Old DDC Numbers

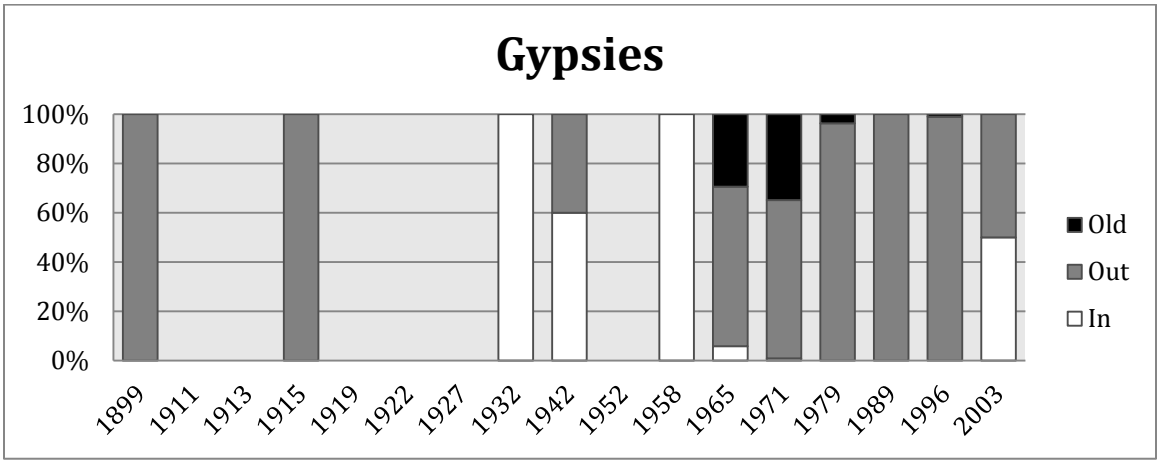

Figure 6. Visualization of the Percentage of Gypsies Books Classed In, Out, and in Old DDC Numbers (NB: years with no data show no bar)

We have to treat with skepticism the percentage of books "Out" of classes possible from 1965 onward because of the use of Tables from that point. However, this exploratory data analysis allows us to reflect on the nature of subject in the DDC and following Langridge's analysis a clearly understanding of the ramifications of forms of knowledge and topics as analyzed in subject analysis and representation.

Langridge

In his 1989 work *Subject Analysis: Principles and Procedures* Langridge outlines an analytical rubric for interpreting a text for its subject matter. He specifically addresses what he sees as the confusion between forms of knowledge, like PHILOSOPHY or NATURAL SCIENCE, and topics, such as HORSES. He outlines a robust set of interpretation guides for the cataloguer and indexer. His work is useful, but is based on anecdotal evidence. With the data we are collecting using the Z39.50 protocol we can begin to lay data next to his work.

In this case we see topics EUGENICS ANATOMY and GYPSIES each with varying degrees of valence. That is, ANATOMY seems to demonstrate a strong valence with art and applied science. We do not get as strong a valence with the other two topics throughout their ontogeny. We see a kind of ambivalence with EUGENICS and GYPSIES. Further, these latter two topics are different in kind. One is a kind of research and practice. The other is a group of people. Perhaps we need to consider these topics differently in the context of a forms of knowledge classification scheme?

Emerging research on the treatment of people in classification schemes has problematized the position of different groups, and here we see the ramifications of scheme change on the positioning of GYPSIES And as with EUGENICS we lose the historical context of the term in a long-lived collection based on updates to the semantics to reflect our contemporary vision of people and science. Given this, perhaps we can extend Langridge's atemporal conception of subject analysis to account for this time-

sensitive valence of topics to forms of knowledge, and begin to craft policy, practice, and technological innovations to classification work.

References

Langridge, D. W. (1989). Subject analysis: principles and procedures. Bowker.

Miksa, F. (1998). The DDC, the universe of knowledge, and the post-modern library. Albany, NY: Forest Press.

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