First Level: Main Headings Only

It is clear that most of the “skew” in the top level of the schema comes from just two main headings: Business, Economics and Finance and Places. These two have delta values of 43.2% and 26.77%, respectively, while the other 14 headings have values ranging from 3.68% to 0.16%. In fact, the other 14 categories’ delta values combined equal less than that of “Places” (19.45% vs. 26.77%).

Business, Economics and Finance is the most obvious outlier, accounting for over half of all occurrences of UNTL-BS terms yet comprising only 8% of the narrower terms in the schema.

The second most obvious outlier, Places, is skewed in the opposite way: it comprises almost 45% of the narrower terms in the schema yet accounts for only 18% of the occurrences. (At first glace this seems like it will be perfectly justified, since place names will understandably get very specific in a collection that centers on local history).

The only other main heading that is skewed toward having more occurrences than narrower terms is Architecture, although this ratio is much less drastic: it accounts for 2.7% of all UNTL-BS occurrences yet comprises only 1.4% of the narrower terms.

All of the other branches of the UNTL-BS schema are lop-sided in the direction of having more narrower terms than would seem to “match” their number of occurrences.

Second Level: One Narrower Term

While, on the level of main headings, it appears that most of the branches are relatively well-balanced, examining one layer further down reveals that, in many cases, this balance only appears in aggregate, as the result of sub-branches that skew in opposite directions cancelling each other out.

1. BEF has a range of delta values within its branch. The top three most-used narrower terms (“Communications,” “Advertising,” and “Journalism”) have delta values of 27-32% each, which are the result of extraordinarily high occurrence counts relative to the number of narrower terms beneath them. The next three most-used terms (“Transportation,” “Oil and Gas,” and “Stores”) are in the opposite situation: each has a delta value between 15 and 17%, which are the result of having a high number of narrower terms relative to their occurrence counts. After these six most-used terms, the delta values drop to 7% and lower. Finally, the main heading alone is the sixth most well-used of the 24 narrower terms on this branch, which means that this branch is definitely top-heavy.
2. Places is lop-sided within the overall structure, but it appears to be fairly well-balanced within its branch. The narrower term “United States” is definitely the most central, comprising 99% of the occurrences and 92% of the narrower terms. But this amounts to a delta value of only 6% on this narrower term, which is lower than the outliers I identified on BEF. The main heading alone is the second to least-used on the branch, which means that this branch is not top-heavy.
3. In Agriculture, “Domestic Animals” has a high occurrence count relative to its number of narrower terms, and “Farming” has a high number of narrower terms relative to its occurrence count. The main heading is fifth in the list of seven narrower terms – not bad, but still perhaps a bit top-heavy.
4. Sports and Recreation is lop-sided primarily due to a single: “Football,” which has a very large number of occurrences for its number of narrower terms. Also, the main heading alone is the third most-used term on the list, accounting for almost 9% of the total occurrences of this branch within the portal.
5. Social Life and Customs has four narrower terms with high delta values. The first two are “Correspondence” and the main heading alone, which both have high occurrence rates for their number of narrower terms. The last two are “Clubs and Organizations” and “Customs,” which both have a higher number of narrower terms than is warranted by their number of occurrences. Furthermore (as is expected by its high delta value), the main heading is very high on the list – third, to be exact, with nearly 15% of the total occurrences on the branch. This indicates that this branch is very “top-heavy.”
6. Education has two extreme outliers: “Schools” and “Colleges and Universities.” The former has a very high occurrence count (over 30% of the total occurrences on the branch), but a relatively low number of narrower terms (11% of the total). The latter has an extraordinarily high occurrence rate *and* narrower term count, but the narrower term count is still considerably greater than it should be (this term carries over 50% of the occurrences in the branch, but over 75% of the narrower terms). This seems logical given that I assume these narrower terms are the names of schools. The main heading branch is relatively high on the list of narrower terms (it is the fourth most-used), which means that the branch is also top-heavy.
7. Religion has two major outliers: “Missions” and “Denominations.” “Missions” has a very high number of narrower terms for having a very low occurrence count, while denominations has high values in both, but a slightly higher occurrence count. The main heading alone is the third most-used term on the list, which means that this term is also top-heavy.
8. Military and War has one major outlier – “Personnel” – and one minor – “Forts.” The former has a very high occurrent count for having only two narrower terms, while the latter has a low occurrence count for having four narrower terms. The main heading is near the middle in terms of usage (7th out of 11), but its usage count is pretty similar to the whole bottom half of the list, so I don’t think I can call this branch top-heavy.
9. People has two *extremely* lop-sided terms. “Ethnic Groups” accounts for 20% of the occurrences from the branch, yet contains two-thirds of the narrower terms (this may be largely explained by the fact that many of the narrower terms under it are the names of native tribes, which are present in the Gateway to Oklahoma History, but not in the Portal. The other lop-sided term is “Individuals,” which accounts for over half of the branch’s occurrences in the Portal, but contains only one-quarter of its narrower terms. The main heading alone is also the fourth most-used out of the seven narrower terms on the list, which means that the branch is pretty top-heavy. This is especially suspicious to me since describing an item as containing “people” doesn’t seem like a particularly valuable subject.
10. Architecture has three major outliers: “Buildings,” “Civil Works,” and “Landmarks.” The first has way too many occurrences for the number of narrower terms it has (almost three-quarters of the occurrences for having only one – i.e. no – narrower terms). The other two have a large number of narrower terms relative to their number of occurrences. The main heading is also the fourth out of eight most-used terms, which makes the branch top-heavy as well.
11. Landscape and Nature has three major outliers. “Geography and Maps” accounts for over 60% of the total occurrences in the branch, yet it has only one – i.e. no – narrower terms. By contrast, “Water” and “Natural Disasters” each account for only around 8% of the total occurrences on the branch, yet they each account for 25-30% of the total number of narrower terms. The main heading alone falls at around the middle of the terms list as far as occurrences, go, with about 2.5% of the total. This does make the branch a little top-heavy, but as this occurrence count is not too far off from the other terms that have no narrower terms, it’s not too egregious.
12. Government and Law has only one real outlier: “Texas Laws and Regulations.” This is to be expected given the focus of the Portal’s collections, so the fact that this term accounts for 13% of the total occurrences on the branch while having no narrower terms under it is understandable. And, given that there are many other sibling terms that occupy this level of the branch with “Texas Laws and Regulations,” and that these are all more-or-less even in terms of occurrence and spread, I would say that this branch is actually pretty well-balanced. However, the main heading alone does fall high on the list: 6th out of 21. Although its occurrence count does not far exceed those of its siblings, in this case the fact that it has so many siblings may suggest that the branch is top-heavy (i.e., if there has been so much attention given to adding narrower terms on this branch, why are there still so many records that are using the generic term instead of one of them?)
13. Literature’s extreme outlier is the main heading itself, which accounts for a whopping 62% of the total occurrences in the branch, while its four siblings account for 17% (“Poetry”), 8% (“Fiction”), 6% (“Children’s”), and 5% (“Folklore”), respectively. Since none of them have any narrower terms, we can’t necessarily speak of this branch in terms of spread beyond the second level. However, the fact that so many records contain the term “Literature” but not a further subdivision when such subdivisions already exist for some genres raises some questions. Is it that these records actually contain types of literature that wouldn’t be well-suited by the use of the existing four sibling terms? Or is it just that one of these siblings would have been appropriate, and the generic term was applied instead in error?
14. Science and Technology is in much the same situation as Literature: the single extreme outlier is the main heading, which accounts for two-thirds of the occurrences in the branch. In this case, I know exactly why this is, since all or most of the items in the Texas Patents collection have been given this term.
15. Arts and Crafts has two extreme outliers and a third that is more minor. The greatest outlier, “Crafts,” occurs relatively sparsely in the Portal, accounting for only 2-3% of the total occurrences on the branch, yet it contains 44% of the branch’s narrower terms. The second outlier, “Music,” accounts for over half of the occurrences on the branch and yet contains less than a quarter of the branch’s narrower terms. The less extreme outlier is “Theatre,” which accounts for 19% of the occurrences on the branch yet has only 8% of the branch’s narrower terms. The main heading is also the third most-used out of the eight terms on the list, accounting for over 9% of the occurrences. Like in the case of Literature, this begs the question of whether there are other types of Arts and Crafts that are prevalent in the Portal but not represented in the browse structure, or whether the sibling terms already in existence would be appropriate for many of these records, and the generic term was simply mis-applied.
16. Finally, Immigration has only two items on its list. The first is the main heading, which accounts for 60% of the occurrences, and the second is its single narrower term, “Colonies,” which accounts for 40%. Given that the total occurrence count of this branch as a whole is very low (less than 1,500), I don’t think that adding any more terms is likely to be warranted.

There are 38 narrower terms with 10%+ delta values.

* Exactly half (19) of them comprise a higher percentage of their branch’s total narrower terms than they do percentage of their branch’s total occurrences. The question to ask of these terms are 1) whether all of their narrower terms are providing enough benefit to users and/or serving a clear enough need in the collections to remain in use, and 2) if there are any narrower terms at lower levels in the branch that are particularly responsible for causing this delta value in the higher level (that is, is the second level off-balance because all of the lower levels are, or because only one or a handful are?)
* The remaining 19 comprise a higher proportion of their branch’s occurrences than its narrower terms. The questions to ask of these terms are 1) has the more general term been applied meaningfully to all of these records, and 2) are there are any topics in the branch that might call for an additional narrower term?

Third Level: Two Narrower Terms