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

The 19 second-level narrower terms that comprise a higher percentage of their branch’s narrower terms than of its total occurrences are:

1. Agriculture – Farming

* Many of the narrower terms in this branch are not used very much. For example, five are used less than 100 times: Rice, Pecans, Potatoes, Peanuts, and Soybeans. This suggests hat perhaps several of these narrower terms are not really necessary.
* However, it is also true that the main heading and single narrower term by themselves constitute the largest percentage of the occurrences (31.8%). This, combined with the fact that it has 16 further narrower terms, many of which are hardly used, may mean that there could be new, more beneficial topics under farming. This is something that clustering or topic modeling might be able to identify.

1. Architecture – Civil Works

* This branch has only one narrower term: “Dams,” which is under the term “Locks and Dams.” On first glance, this seems pretty obviously unnecessary, as the items labeled “Dams” could likely equally be served by the more generic term. If, however, there are enough dams represented in the collections that don’t include locks, then I would say that the generic term should be changed to “Dams” with “Locks” as a narrower term – since you wouldn’t have locks without a dam, but you can have dams without locks.

1. Architecture – Landmarks

* This branch, to my eyes, looks reasonable. None of the three narrower terms under “Landmarks” has any further narrower terms, and the term “Historic Markers” seems justified in being used a great deal, since there are many different markers to appear in the collections. It may be possible that one or more of the marker sites would be justified in having its own narrower term - similar to the Alamo – which is something that clustering or topic modeling may help to identify. If such an additional term was added, this might help with the only possible case of overuse of the “Historic Markers” term I can think if – i.e. if people have ascribed the term for the markers themselves to the sites to which they are attached.

1. Arts and Crafts – Crafts

* This branch has one term with several additional terms underneath it. That term is “Fiber Arts,” and the further narrower terms are Knitting, Sewing, Spinning, and Weaving. Knitting and Spinning seem like they may not be justified (they are only used 26 and 39 times, respectively), and weaving isn’t used that much either (126 times). However, Sewing is very prolific, being used over 1,400 times. Given this, we can conclude that the low delta score on this branch is misleading: although “Fiber Arts” comprises 45% of the narrower terms and 47% of the occurrence count, about ¾ of this count comes from a single term, and the others are really pretty weak. If future analysis (such as clustering or topic modeling) suggests that there is no good narrower term for it, then it’s possible that the section should be simplified and reduced down to just “Fabric Arts” across the board.
* Stained Glass may not be a great term, either, since it is used less than 100 times (93, to be exact).

1. Business, Oil, and Finance – Oil and Gas

* “Companies” has 7 narrower terms in this branch and “Oil Fields” has 2.
* “Companies” also carries the plurality of the branch’s occurrences (31%) and thus perhaps justifies its many narrower terms. However, on closer inspection, two of these narrower terms aren’t pulling their weight. One, “Exxon” makes sense since it is not an authorized UNTL-BS term and will eventually be removed from the record it is on anyway. But the second, “Phillips 66,” is an authorized term, yet is used only 20 times. This suggests that it may likely be more suited to being a keyword than an authorized browse term.
* “Oil Fields” is used only 285 times. Out of these 285 uses, 251 are the generic terms – which argues powerfully in favor of demoting its two narrower terms (“Ranger Oil Field” and “Spindletop”) to custom keywords.

1. Business, Economics and Finance – Stores

* This branch is dominated by its generic term – which is strange, considering that it has 13 narrower terms to choose from. This means that either 1) the existing narrower terms are in fact appropriate for some of the records that instead use the generic term, or 2) there are additional types of stores that aren’t yet included in the schema but that should be.
* In case #1,

1. Business, Economics and Finance – Transportation

* “Automobiles”, which is the most-used term on this branch, gets no narrower terms while “Aviation” and “Railroads” both get 2 and “Trucks” gets 1. It may be that this is not a problem, since there are two major categories under aviation and a major distinction between personal trucks and commercial trucks, but no such categorical differences to be laid out under cars. However, it may be enlightening to do some clustering or topic modeling on this narrower term nevertheless to see if any obvious potential sub-terms rise to the surface.
* “Airplanes” and “Helicopters” both pull their weight under “Aviation” – the former is responsible for more occurrences than the generic term, and the latter boasts about half as many as the generic term.
* “Commercial” under “Trucks” isn’t as clearly useful, but it does enjoy one occurrence for every 2-3 on the generic term.
* One last small nitpick that probably doesn’t deserve attention would be that “Horse-Drawn Vehicles” should be a narrower term under “Animal-Drawn Vehicles”.

1. Education – Colleges and Universities

* This branch is full of a large number of very specific narrower terms. This is with good reason, as each narrower term represents a single institution. However, there are many which are not used very frequently, and might therefore do better as keywords. These include twelve which are used fewer than 100 times: Austin College, Palo Alto College, El Centro College, North Lake College, Richland College, Eastfield College, Mountain View College, Cedar Valley College, Clarendon College, Brookhaven College, Southwestern Christian College, and St. Mary's University.
* There is a large number of records that contain just the generic term – it would be beneficial to use clustering or topic modeling to check for any new potential terms that particularly stand out. In this case it may also be possible to look at the range of values in a few of the records’ fields (such as author, collection, or title) to figure out if any additional school names appear in high enough numbers to justify a browse term.
* Unsurprisingly, the University of North Texas has a hugely disproportionate occurrence count. For the most part, this is likely justified.
* “Faculty and Staff”, however, should probably be moved to be a first-level narrower term – a sibling of “Colleges and Universities” and not its child. This is justified both conceptually (as there are faculty and staff in other types of education institutions) and numerically, by the over 32,000 occurrences of the term. Whether the term gets moved or not, it could be useful to try clustering or topic modeling to find suggestions for sub-terms to siphon off some of those occurrences.

1. Immigration – Colonies

* This branch and its parent term make an interesting case. “Colonies” is the only sub-term below “Immigration”, and it can definitely be said that it carries its weight, comprising almost 65% as many occurrences as its parent term. Numerically, then, the branch is very well-balanced. Yet I wonder whether the term “Colonies” is being used in the same sense every time it is added to a record in the collection. If, in fact, it is being used in multiple senses, then this may compromise the apparent balance in the branch.

1. Landscape and Nature – Natural Disasters

* This term is actually fairly balanced. Unless there is some really meaningful categories of fire, the fact that “Fires” has a high occurrence count without any corresponding narrower terms is understandable as a simple reflection of the contents of the collections.
* The rest of the terms seem balanced as well – with the sole exception of “Dust Storms”, which is used only 43 times. Unless there is an important subjective reason to include “Dust Storms” on the same level as more numerous types of natural disaster, it appears that this term would be better served as a keyword.

1. Landscape and Nature – Water

* This term appears at first glance to be one where the generic term was simply misapplied. One struggles to understand what kinds of natural water the records with the generic term could be referring to if not any of the narrower terms already provided. Clustering or topic modeling might be able to confirm this – although, once it is confirmed there isn’t really much we can do to rectify it manually.

1. Literature – Children’s

* Three of the four narrower terms under “Literature” returned high delta scores only because the branch as a whole doesn’t have many occurrences. The problem is more that the generic term has an oddly-high number of occurrences – possible remedies for which are explained below, under that term.

1. Literature – Fiction

* Three of the four narrower terms under “Literature” returned high delta scores only because the branch as a whole doesn’t have many occurrences. The problem is more that the generic term has an oddly-high number of occurrences – possible remedies for which are explained below, under that term.

1. Literature – Folklore

* Three of the four narrower terms under “Literature” returned high delta scores only because the branch as a whole doesn’t have many occurrences. The problem is more that the generic term has an oddly-high number of occurrences – possible remedies for which are explained below, under that term.

1. People – Ethnic Groups
2. Religion – Missions
3. Science and Technology – Chemical Plants
4. Social Life and Customs – Clubs and Organizations
5. Social Life and Customs - Customs

The 19 second-level narrower terms that comprise a higher percentage of their branch’s total occurrences than of its narrower terms are:

1. Agriculture – Domestic Animals:

* The main heading and single narrower term by themselves constitute the largest percentage of the occurrences (34.5%). Although it does have 10 narrower terms, the fact that so many records still use the generic terms may mean that there are one or a few other types of domestic animals that deserve a narrower term. Perhaps clustering or topic modeling would help identify them. The other possibility is that this term was frequently mis-applied, and that some of the existing narrower terms could have been used on many of these records instead.
* “Cattle” comprises over 25% of occurrences yet has no narrower terms. This seems reasonable, since we can reasonably expect that an uncommonly large percentage of the collections’ materials has to do with cattle. The only exception would be if there is some sort of special cattle that are numerous enough in the collections to warrant their own term.
* “Poultry” has one further narrower term, yet it is not being used that much (“Poultry by itself occurs 719 times, and “Turkey” 154 times). This may be a narrower term that doesn’t need to exist, unless there’s a compelling research reason to single out turkeys.
* “Horses” has one further narrower term. Its delta suggests that perhaps it should have more than one narrower term, since it occurs with great frequency in the collections.

1. Architecture – Buildings

* This branch ends with the first narrower term, yet it occurs over 100,000 times in the collections. This indicates that either 1) the branch desperately needs more narrower terms to make it less top-heavy, or 2) the generic term is being over-used on records where it is not really helpful. In case #1, clustering or topic modeling may help suggest possible narrower terms. On the other hand, examining the range of metadata values in a few select fields (such as collection and resource type) may help prove or disprove the hypothesis of case #2.

1. Arts and Crafts – Music

* This branch is dominated by its generic term, which monopolizes 84% of its occurrences. A clustering or topic modeling exercise might suggest if there are any better narrower terms to add to the branch, since none of the existing narrower terms seems scrawny enough to be replaced.

1. Arts and Crafts – Theatre

* This branch has only one narrower term, which siphons off a respectable portion of the total occurrence count, but that still leaves a massive piece that appears to need additional narrower terms. More analysis needs to be done in order to determine in what context this more generic term was used throughout the collections (for instance, was it used to refer to specific shows, specific buildings, or perhaps materials related to the art of theatre). Only after determining this can additional narrower terms be chosen.

1. Business, Economics and Finance – Advertising

* This branch is the second-most-used in the entire schema, yet it has no narrower terms at all. First, checking the range of metadata in some key fields (e.g. collection and resource type) may help determine whether some of this high occurrence count has been improperly added. Second, doing a clustering or topic modeling exercise may help suggest useful narrower terms.

1. Business, Economics and Finance – Communications

* This branch, unlike so many others I’m investigating doesn’t suffer from having a lot of occurrences on its generic term. Rather, it is *extremely* off-balance in favor of, primarily, “Newspapers,” but also secondarily “Media.” Both of these have hundreds of thousands of uses in the collections and no narrower terms at all. Like with “Advertising,” the first order of business will be to try and determine if all 858,000 “Newspapers” and 244,000 “Media” have been accurately applied, or if many of them have been applied en masse and should be removed. Then, using clustering or topic modeling may suggest appropriate narrower terms based on the occurrences that remain after the branch is culled.

1. Business, Economics and Finance – Journalism

* As with “Advertising” and” Communications – Newspapers”, this branch has over 800,000 uses in the collections and no narrower terms at all. As with the above examples, I will need to perform a metadata range analysis to try and determine which of the occurrences are accurate or helpful, and then use clustering or topic modeling to try and suggest meaningful narrower terms to siphon off some of the remaining usages.

1. Education – Schools

* This branch is dominated by the term “Students”, which comprises over 66% of the occurrence count yet has no narrower terms. This could be because it has been applied inappropriately – in which case examining the range of metadata values in the collection or resource type fields may help determine this. Or, it may be that the occurrences are legitimate, in which case clustering or topic modeling may suggest potential narrower terms.
* As a side-note, it appears that “Faculty and Staff” under “Colleges and Universities” may be better served being under this branch of the schema, which seems to be devoted more toward types of people where the former is otherwise entirely dedicated to specific institutions.

1. Government and Law – Texas Laws and Regulations

* This branch comprises over 32,000 occurrences, yet has no narrower terms at all. This is less egregious than “Advertising”, “Communications – Newspapers”, and “Media”, but it still may be helpful to run a clustering or topic modeling exercise on its records to see if there are any potential narrower terms that stand out.

1. Immigration [main heading only]

* This term is, in reality, not very off-balance at all – unless the sense of its one narrower term, “Colonies”, is being misinterpreted across the collections.

1. Landscape and Nature – Geography and Maps

* This branch has over 85,000 occurrences and no narrower terms. This means that either it has been applied too liberally (which a metadata range analysis will help to determine) or that it needs one or more narrower terms (which clustering or topic modeling might be able to suggest).

1. Literature [main heading only]

* This whole branch is not very well-used. It has four narrower terms, which all see just enough usage to seem worthwhile (there is also the benefit of subject judgment, which tends to support keeping the narrower terms even if they are underused, because of the strong semantic distinction among different types of literature).
* That being said, it is also true that this generic term has far more occurrences than I would expect in a case where four narrower terms exist. This suggests that either one of the existing terms would be more appropriate for many of the records that use the generic term, or that there are one or more main divisions of literature that could be added.

1. Military and War – Personnel
2. People – Individuals
3. Religion – Denominations
4. Science and Technology [main heading only]
5. Social Life and Customs – Correspondence
6. Social Life and Customs [main heading only]
7. Sports and Recreation - Football