UNTL-BS Spread Vs. Usage Analysis

1. Business, Economics and Finance **51.47%** of occurrences **8.27%** of narrower terms **43.20%** delta
2. Main heading has too many occurrences

* The main heading alone is used 14,042 times, which makes it the sixth-most-used term among all its branches, after “Communications,” “Advertising,” “Journalism,” “Transportation,” and “Medicine.” However, it comprises only 0.47% of the occurrences on the heading (because “Communications,” “Advertising,” and “Journalism” are such extreme outliers), and only 0.24% of the occurrences on the Portal as a whole.
* Perform metadata variability analysis to determine whether many or most of these occurrences come from one or a few collections
* Perform topic modeling to determine if there are narrower terms that could be added that would compare to the existing ones for usage.

1. ~~“Transportation” has too many narrower terms~~
2. ~~“Oil and Gas” has too many narrower terms~~
3. ~~“Stores” has too many narrower terms~~
4. “Communications” has too few narrower terms

* The terms on this branch are used over one million times (1,106,557), meaning that this branch is responsible for 19.12% of all UNTL-BS occurrences on the Portal! Yet it has only 3 narrower terms: “Newspapers,” “Media,” and “Telephones.”
* The main heading + first narrower term doesn’t appear very often, which means there is little benefit to performing more analysis on this level.
* On the third level (main heading + 2 narrower terms), we see that 77.62% of the branch’s occurrences come from the narrower term “Newspapers,” which itself has no narrower terms. Perform metadata variability analysis on this second narrower term to determine whether many or most of its occurrences come from one or a few collections, then perform topic modeling to determine if there are narrower terms that could be added.

1. “Advertising” has too few narrower terms

* This branch actually has *no* narrower terms, yet it comprises 14.76% of *all* occurrences on the Portal! (853,926 out of 5,786,278). This means that it is either in *desperate* need of narrower terms to make it useful for browsing, or it has been drastically over-applied.
* Perform metadata variability analysis to determine the extent of its over-use (and whether it comes mainly from one or a few collections)
* Perform topic modeling to suggest possible narrower terms.

1. “Journalism” has too few narrower terms

* This branch actually has *no* narrower terms, yet comprises 14.66% of *all* the occurrences on the Portal! (848,300 out of 5,786,278). This means that it is either in *desperate* need of narrower terms to make it useful for browsing, or it has been drastically over-applied.
* Perform metadata variability analysis to determine the extent of its over-use (and whether it comes mainly from one or a few collections)
* Perform topic modeling to suggest possible narrower terms.

1. Places **18.12%** of occurrences **44.89%** of narrower terms **26.77%** delta
2. “United States” is clearly the dominant branch, but it is balanced (since almost all narrower terms pertain to Texas, but the Texas terms comprise almost all of the occurrences)
3. Social Life and Customs **4.93%** of occurrences **7.34%** of narrower terms **2.41%** delta
4. Main heading has too many occurrences

* The main heading alone is used 42,125 times, which makes it the fourth-most-used term among all its branches, after “Correspondence,” “Customs,” and “Clubs and Organizations.” It comprises 14.76% of the occurrences on the heading, which corresponds to only 0.73% of the occurrences on the Portal as a whole.
* Perform metadata variability analysis to determine whether many or most of these occurrences come from one or a few collections
* Perform topic modeling to determine if there are narrower terms that could be added that would compare to the existing ones for usage.

1. ~~“Customs” has too many narrower terms~~
2. ~~“Clubs and Organizations” has too many narrower terms~~
3. “Correspondence” has too few narrower terms

* This branch is used 475,528 times, comprising 16.65% of all occurrences under its main heading, yet it has no narrower terms.
* The high occurrence rate of this term may be due to its mis-use: to merely identify the record as a letter, which would be a duplication of the record type field.
* Perform metadata variability analysis to determine whether many or most of these occurrences come from one or a few collections
* Perform topic modeling to determine if there are narrower terms that could be added.

1. People **4.92%** of occurrences **6.32%** of narrower terms **1.40%** delta
2. “Individuals” has too few narrower terms

* Despite the fact that this branch has 16 narrower terms, its 158,187 occurrences would seem to suggest that it needs more.
* On closer inspection, it is clear that its existing narrower terms are extremely under-used given the number of occurrences on the branch as a whole. Over 99% of occurrences come from the main heading + first narrower term, and many of the narrower terms are used only a handful of times.
* This would seem to be a prime candidate for mis-use since, like “Clothing,” it may have been applied to records simply when an individual is depicted in the resource, as opposed to its intended use: when one of a select few individuals is present or referred to in the record.
* Perform topic modeling to identify any proper names that may warrant a narrower term.

1. ~~“Ethnic Groups” has too many narrower terms~~
2. Government and Law **4.27%** of occurrences **4.74%** of narrower terms **0.47%** delta
3. “Texas Laws and Regulations” has too few narrower terms

* This branch is used 32,840 times, comprising 13.28% of all occurrences under its main heading, yet it has no narrower terms.
* Perform topic modeling to suggest possible narrower terms.

1. Education **3.91%** of occurrences **5.86%** of narrower terms **1.95%** delta
2. “Schools” has too few narrower terms

* Despite the fact that this branch has 6 narrower terms, its 73,988 occurrences would seem to suggest that it needs more.
* Within this branch, “Students” comprises the majority of occurrences (66.69%). Perform metadata variability analysis to determine whether many or most of these occurrences come from one or a few collections, and topic modeling to suggest possible narrower terms.

1. ~~“Colleges and Universities” has too many narrower terms~~
2. Arts and Crafts **2.52%** of occurrences **2.32%** of narrower terms **0.2%** delta
3. ~~“Crafts” has too many narrower terms~~
4. ~~“Music” has too many narrower terms~~
5. “Theatre” has too few narrower terms

* This branch and its single narrower term – “Motion Pictures” – are used 27,851 times, comprising 19.08% of the occurrences under its main heading.
* While “Motion Pictures” definitely carries its weight (almost a quarter of the occurrences on the branch), that still leaves over 75% of occurrences on the main heading + first narrower term. Perform metadata variability analysis to determine whether many or most of these occurrences come from one or a few collections, and topic modeling to suggest possible narrower terms.

1. Architecture **2.7%** of occurrences **1.39%** of narrower terms **1.31%** delta
2. “Buildings” has too few narrower terms

* This branch is used 115,444 times, comprising 73.78% of the occurrences under its main heading, yet it has no narrower terms.
* The high occurrence rate of this term may be due to mis-use, if editors applied it merely when buildings are depicted in the resource.
* Perform metadata variability analysis to determine whether many or most of these occurrences come from one or a few collections
* Perform topic modeling to determine if there are narrower terms that could be added.

1. ~~“Civil Works” has too many narrower terms~~
2. ~~“Landmarks” has too many narrower terms~~
3. Landscape and Nature **2.35%** of occurrences **2.88%** of narrower terms **0.54%** delta
4. “Geography and Maps” has too few narrower terms

* This branch is used 85,334 times, comprising 62.88% of the occurrences under its main heading, yet it has no narrower terms.
* Performing metadata variability analysis to determine whether many or most of these occurrences come from one or a few collections may not help prove mis-application in this case, since there are collections on the Portal dedicated to maps which we would expect to be composed mostly of resources that would genuinely receive this heading.
* Perform topic modeling to determine if there are narrower terms that could be added.

1. ~~“Water” has too many narrower terms~~
2. ~~“Natural Disasters” has too many narrower terms~~
3. Military and War **1.57%** of occurrences **3.25%** of narrower terms **1.68%** delta
4. “Personnel” has too few narrower terms

* This branch and its single narrower term – “Veterans” – are used 18,205 times, comprising 20.03% of the occurrences under its main heading (while the branch “Wars” comprises a higher percentage of the main heading’s occurrences, it has a number of narrower terms commensurate with that occurrence rate).
* While “Veterans” definitely carries its weight (22.08% of the occurrences on the branch), that still leaves 77.92% on the main heading + first narrower term. Perform metadata variability analysis to determine whether many or most of these occurrences come from one or a few collections, and topic modeling to suggest possible narrower terms.

1. Sports and Recreation **1.08%** of occurrences **4.28%** of narrower terms **3.2%** delta
2. “Football” has too few narrower terms

* This branch is used 13,165 times, comprising 21.07% of the occurrences under its main heading, yet it has no narrower terms.
* This is less drastic than many of the other examples on this list, because there is a plethora of other branches under this main heading. Unless there is a compelling reason why narrower terms would make semantic sense under “Football,” it may be advisable to leave this branch alone for being relatively well-balanced.

1. Religion **0.86%** of occurrences **2.6%** of narrower terms **1.74%** delta
2. ~~“Missions” has too many narrower terms~~
3. “Denominations” has too few narrower terms

* This branch and its 12 narrower terms are used 29,526 times, comprising 59.18% of the occurrences under its main heading.
* However, very few of these occurrences are from the main heading + first narrower term, which suggests that the way to handle this abundance of occurrences would be to add narrower terms under “Jewish,” “Catholic,” and “Methodist,” which are the three most-used terms on the branch. However, unless there is a compelling reason why adding such terms would make semantic sense, it may be advisable instead to leave this branch alone.

1. Agriculture **0.68%** of occurrences **4.37%** of narrower terms **3.68%** delta
2. “Domestic Animals” has too few narrower terms

* This branch and its 10 narrower terms are used 17,375 times, comprising 43.96% of the occurrences under its main heading.
* Within the branch, the largest percentage of uses come from the main heading + first narrower term, which strongly suggests that there are one or more narrower terms that could be added to make the branch more balanced. Perform topic modeling to determine if there are narrower terms that could be added.
* The fact that the narrower term “Cattle” also comprises a large portion of the occurrences on this branch (25.66%) is understandable given the subject matter of the collections on the Portal. But, it may still be wise to perform some topic modeling to see if there are any narrower terms that could be added.

1. ~~“Farming” has too many narrower terms~~
2. Science and Technology **0.54%** of occurrences **0.84%** of narrower terms **0.3%** delta
3. Main heading has too many occurrences

* The main heading alone is used 20,800 times, which makes it the most-used term among all its branches. It comprises 66.83% of the occurrences under the heading, which corresponds to only 0.35% of the occurrences on the Portal as a whole.
* Perform metadata variability analysis to determine whether many or most of these occurrences come from one or a few collections
* Perform topic modeling to determine if there are narrower terms that could be added that would compare to the existing ones for usage.

1. ~~“Chemical Plants” has too many narrower terms~~
2. Literature **0.05%** of occurrences **0.46%** of narrower terms **0.42%** delta
3. Main heading has too many occurrences

* The main heading alone is used 1,641 times, which makes it the most-used term among all its branches. It comprises 62.54% of the occurrences under the heading – however, this corresponds to only 0.02% of the occurrences on the Portal as a whole.
* Given that the high delta value returned by the main heading is more a result of the small size of the heading overall, it may be advisable to leave this branch alone for being relatively well-balanced.

1. ~~“Folklore” has too many narrower terms~~
2. ~~“Children’s” has too many narrower terms~~
3. ~~“Fiction” has too many narrower terms~~
4. Immigration **0.02%** of occurrences **0.19%** of narrower terms **0.16%** delta
5. Main heading has too many occurrences

* The main heading alone is used 855 times, meaning that it is used more often than its single narrower term – “Colonies.” It comprises 60.60% of the occurrences under the heading.
* Like “Literature” above, the high delta value is the result of the small size of the heading overall, so it would be advisable to leave this branch alone.

1. ~~“Colonies” has too many narrower terms~~

Prime Candidates for Metadata Variability Analysis and Topic Modeling

1. Agriculture – Domestic Animals
2. Architecture – Buildings
3. Arts and Crafts – Theatre
4. Business, Economics and Finance
5. Business, Economics and Finance – Advertising
6. Business, Economics and Finance – Communications - Newspapers
7. Business, Economics and Finance – Journalism
8. Education – Schools
9. Government and Law - Texas Laws and Regulations
10. Landscape and Nature - Geography and Maps
11. Military and War – Personnel
12. People - Individuals
13. Science and Technology
14. Social Life and Customs
15. Social Life and Customs – Correspondence