Collection Evaluation:
IMLS Digital Collections and Content

Spring 2011

This is an evaluation of the content of the integrated IMLS DCC and Opening History aggregations (hereafter DCC). At the time of this evaluation, DCC includes 1,163 publically accessible digital collections and over 1 million items from a broad range of cultural heritage institutions.

The first section of this evaluation briefly summarizes hosting/contributing institutions by type and state; U.S. state coverage of the aggregation as a whole; and major item types in the aggregation.

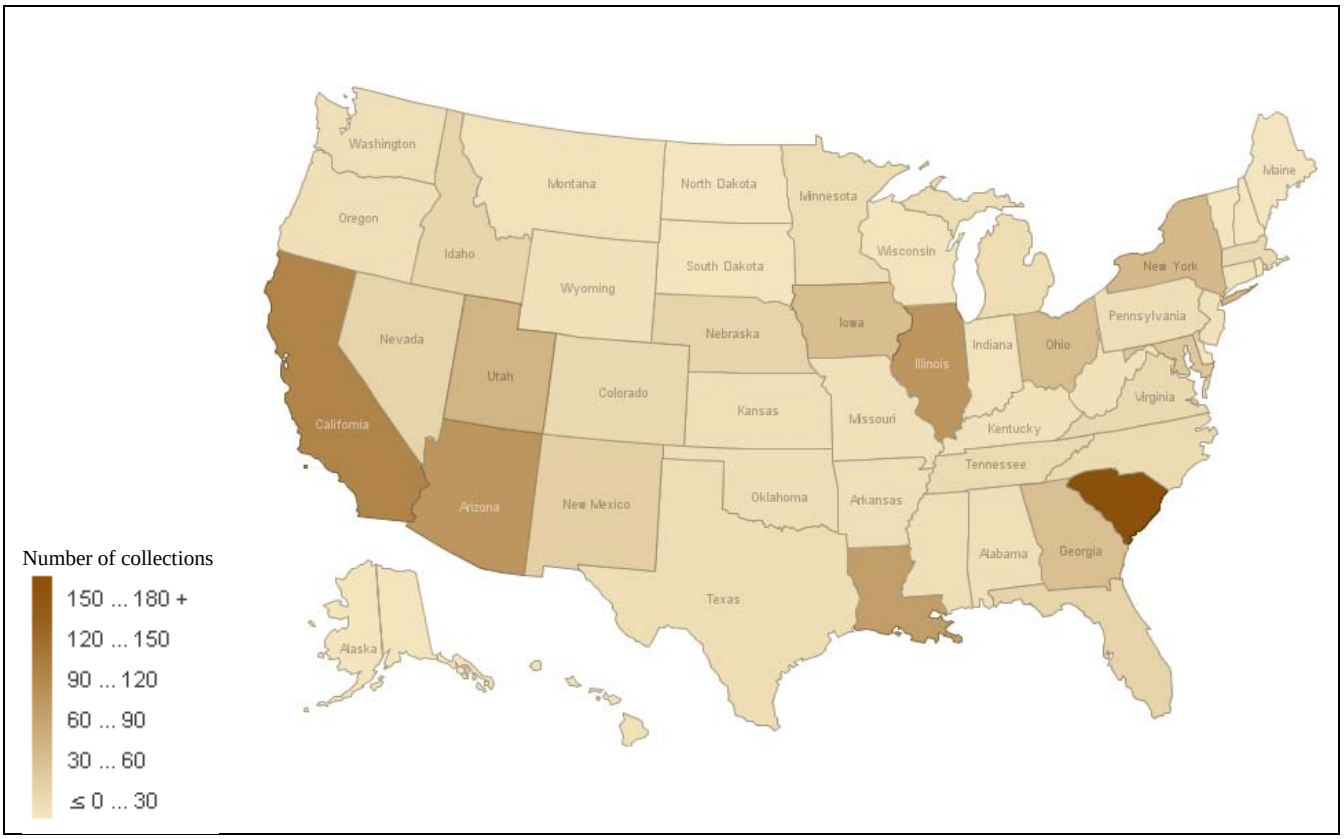
The next section delves into the subject evaluation: method, aggregate-level views, and topic-level views.

The final section describes next steps.

I. Type and Coverage

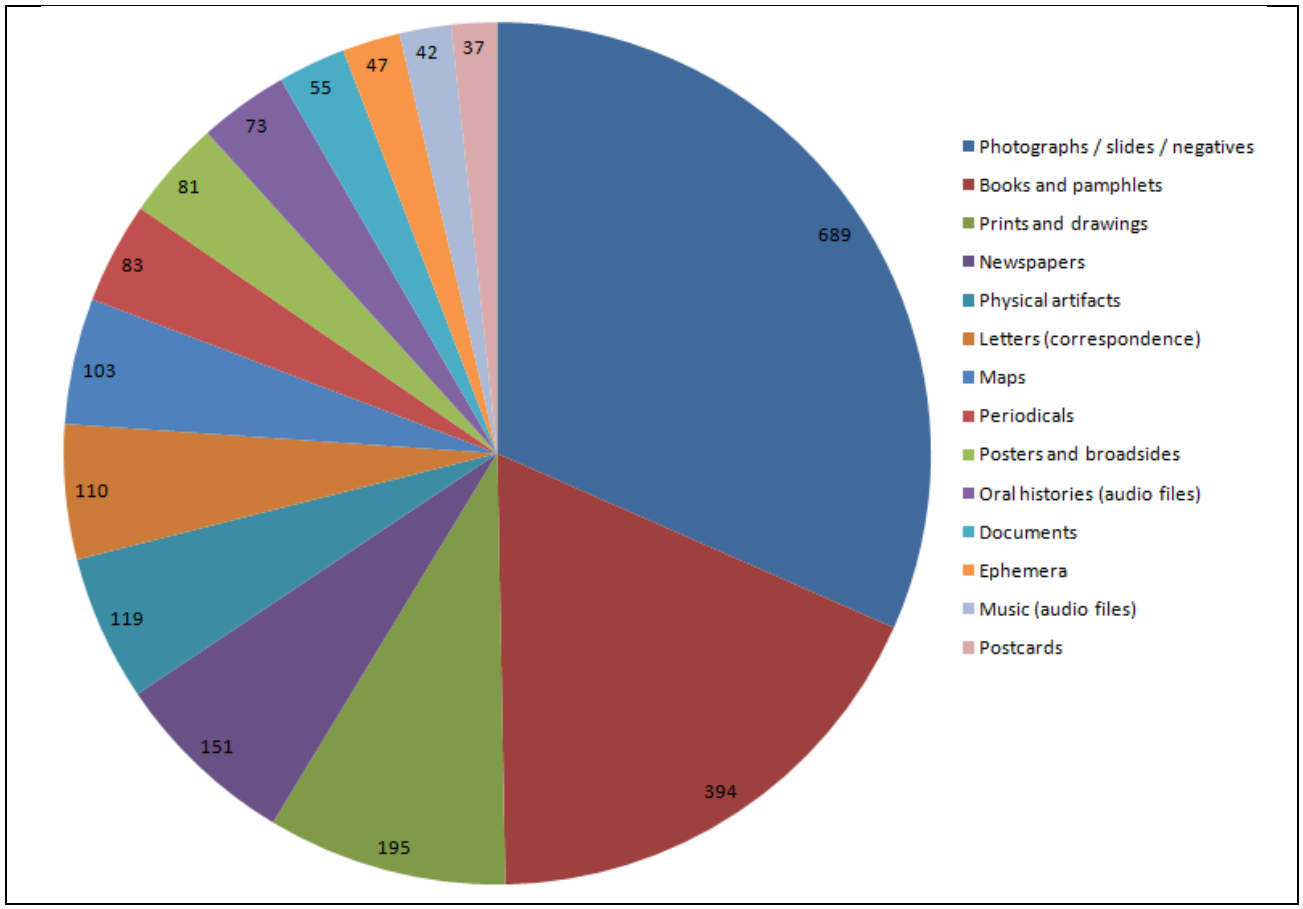
Contributing institutions: Institutions hosting or contributing to collections in DCC range in type and size, and are dispersed in 44 states. Academic libraries contribute about 33% of collections to the aggregation, including 249 collections brought together through the DLF Aquifer initiative. Other major contributors include state libraries, library consortia, public libraries, historical societies and museums. The majority of items in DCC are images, but collections that include texts—primarily books and pamphlets—comprise approximately 40% of the aggregation.

Geographic coverage: While DCC collections currently come from contributing institutions in 44 states, the aggregation offers geographic coverage of all 50 states, with clusters of strong coverage in the South, Midwest, and Southwest.



Heat map of U.S. state coverage in DCC

Item types: The pie chart below represents the diversity of content types in DCC. While image collections dominate the aggregation, collections that include textual objects – such as books and pamphlets, newspapers, and periodicals – constitute about 40% of the aggregation. DCC provides access to many more item types not represented in the pie chart below, including archival finding aids, sheet music, blueprints, lithographs, etchings, field data and field diaries, historical cartoons, etc.



Top 15 item types in DCC as a whole, in terms of number of collections that include each item type

---

2. Subject Evaluation

2a. Method

This evaluation is based on Library of Congress Subject Headings in IMLS DCC and Opening History as of January 18, 2011, just prior to when the number of records jumped to more than 1,000. This information is therefore out of date, but not wildly.

Number of collections at time of data freeze:	864
Collections with 1+ LCSH terms:	691, or 80%

This analysis therefore covers 80% of the collection registry. Analysis of the remaining 20% will follow in the next evaluation.

Evaluation proceeded in this way:

(1) Term by term, I grouped subject headings from the collection registry into overarching categories or topical themes. These categories are informal, vary in their levels abstractness/specificity, and demonstrate significant semantic overlap and tangency. A term can have up to three categories; so far, I have not had to enforce this constraint. Most terms have one or two categories. My identification and assignation of categories prioritized the efficiency of our own, local understanding of our collection registry's contents over scientific rigor. The categories percolate from the data itself. There is little control. There are certainly some flaws, some holes in how I assigned categories to terms. These are priorities for future improvement of our evaluative system.

(2) I also assigned categories subcategories when I discovered sufficient representation of some topic within an existing category. And sometimes, in reverse, subcategories pooled into useful categories over the course of analysis. However, subcategories are not quantitatively treated in the following analysis because they are currently too messy. This is another place for future improvement of our evaluative system: we could, for example, assess subcategory growth to determine at what point a subcategory should be considered a category in its own right. The numerical category codes are arbitrary.

Example of category assignment to term:

Term: "Air bases, American"

	category 1	**subcategory**	**category 2**	**subcategory**	**...**	**...**
Category name	Transportation	aeronautics	War or military history	air bases		
Category code	800		900			

In this case, the term was assigned 2 of a possible 3 categories.

Characterizing the subject data:

Total number of LCSH terms among 864 collections:	2309
Total number of categories assigned to those terms:	36
Total number of subcategories assigned:	223
Uncategorized subject terms*:	731, or 32%

Category	**Code**
Agriculture and ranching	100
Asian American history	200
Athletics	300
Education history	400
Entertainment history	500
Government	600
Religion history	700
Transportation history	800

War or military history	900
African American history	1000
Archeology	1100
Architecture	1200
Biology	1300
Birds	1400
Botany	1500
Civil rights	1600
Ecology, conservation, environmentalism	1700
Emigration and immigration	1800
Hispanic American history	1900
Irish American history	2000
Italian American history	2100
Labor	2300
Literature	2400
Local history	2500
Medicine	2600
Mexican Americans	2700
Mining	2800
Native American history	2900
Oil	3000
Slavery	3100
State history	3200
Women's history	3300
Disasters	3400
Music	3500
Art	3600

*Uncategorized terms include:

- Proper nouns, unless connection to an existing category is readily apparent
- Type and format terms that have been entered as subject terms, e.g. "Diaries" and "Posters"(unless connection to existing topical category is explicit)
- Terms with little distinguishing value, e.g. "Social life and customs" or "United States -- History"
- Most rarely, terms that do not fit into any existing categories and, because of their rarity, don't yet merit creation of new categories, e.g. "Abnormalities, human" and "Snowflakes"

(3) In the following analysis, strength of a category is measured by the number of collections for which a collection-level subject term falls into that category – in other words, the number of collections that include the category. In keeping with our DiCE paper, I distinguish subject-focused from subject-inclusive collections. Subject-focused collections' LSCH terms fall into only one of the categories in my subject scheme. Subject-inclusive collections have terms falling into two or more of the categories in my

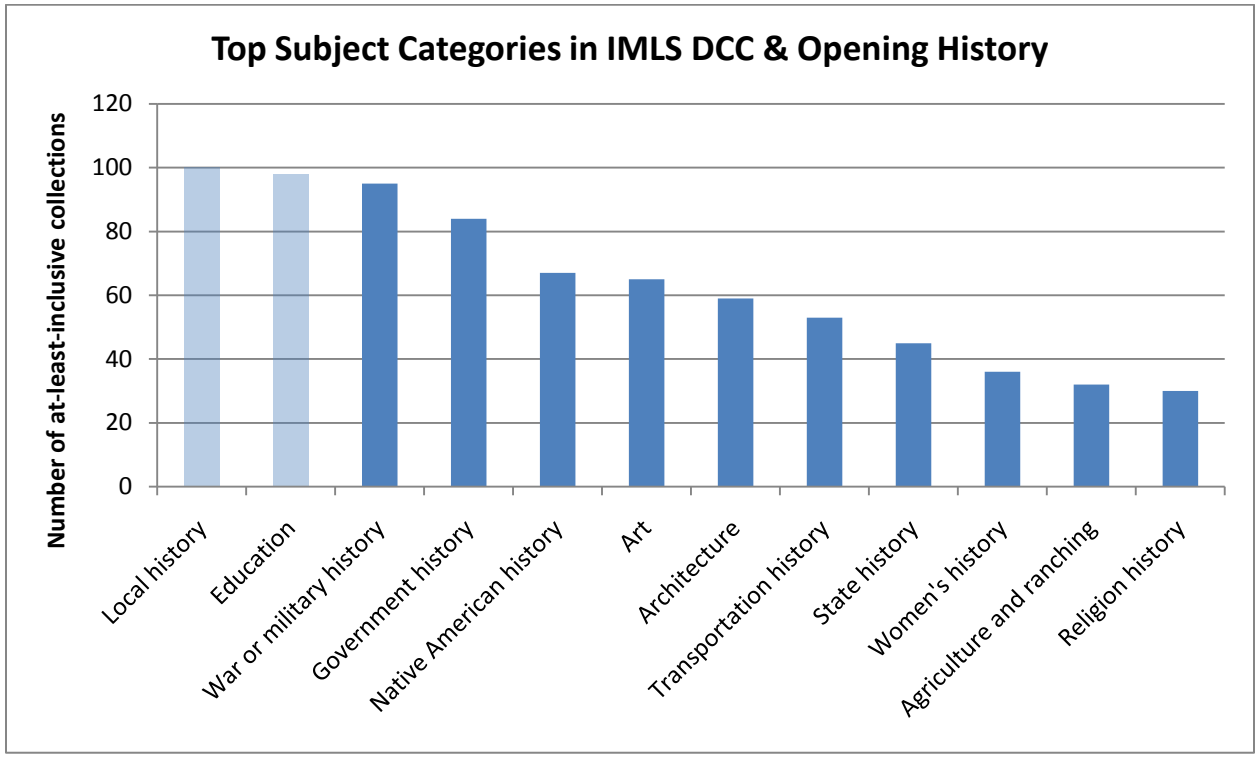
subject scheme. When I refer to “at-least-inclusive” collections, I mean all collections in a category, either inclusive or focused.

(4) In the future, we will add new subject terms from across aggregation, assign category codes (reassigning category codes as needed -- it's easy to make sweeping changes with this approach), and rerun a standard set of queries on a database that relates codes to collection IDs to generate a fresh evaluation.

Significant human attention is still required for this evaluation, though of course the method is intended to be somewhat more mechanical, semiautomatic, repeatable, and efficient that previous evaluations. The initial investment into revising our evaluation method came close to 30 hours of labor. Future evaluations should take a fraction of the time.

2b. Aggregate-level views

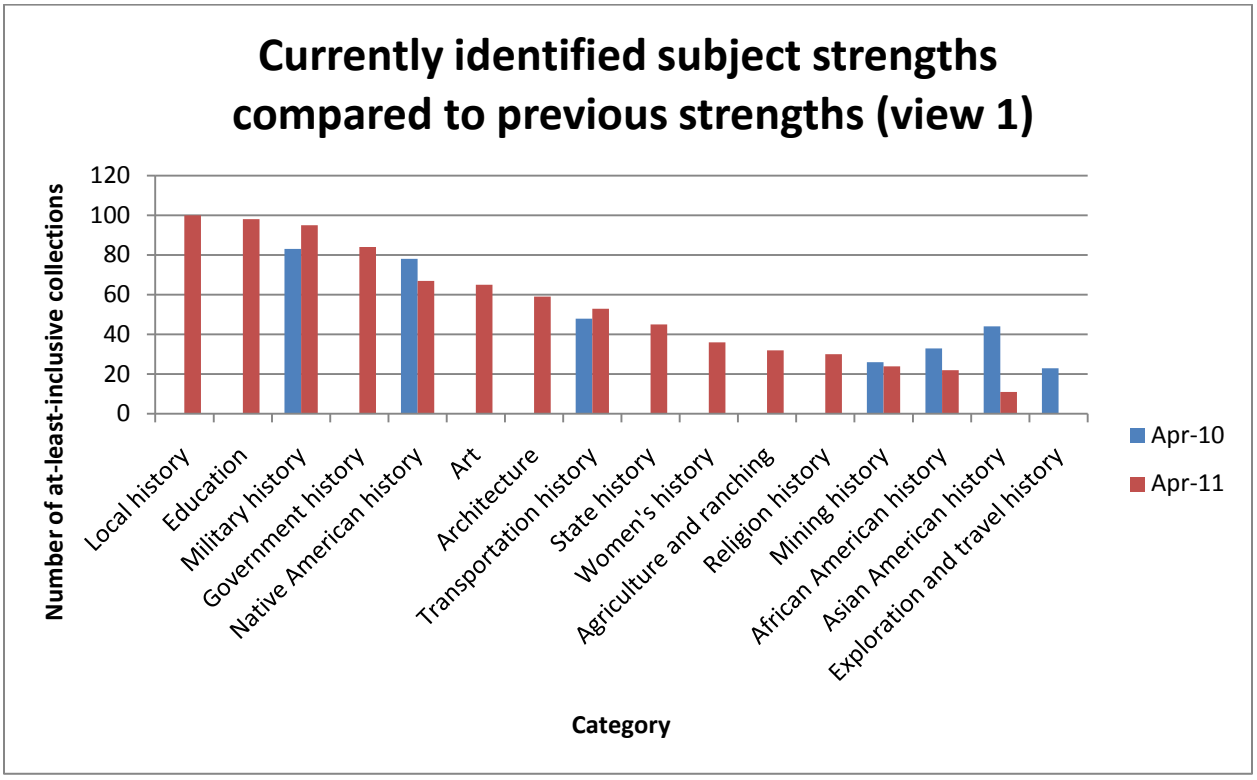
Strictly in terms of numbers of collections that include each category, the following are the top 12 most common categories:



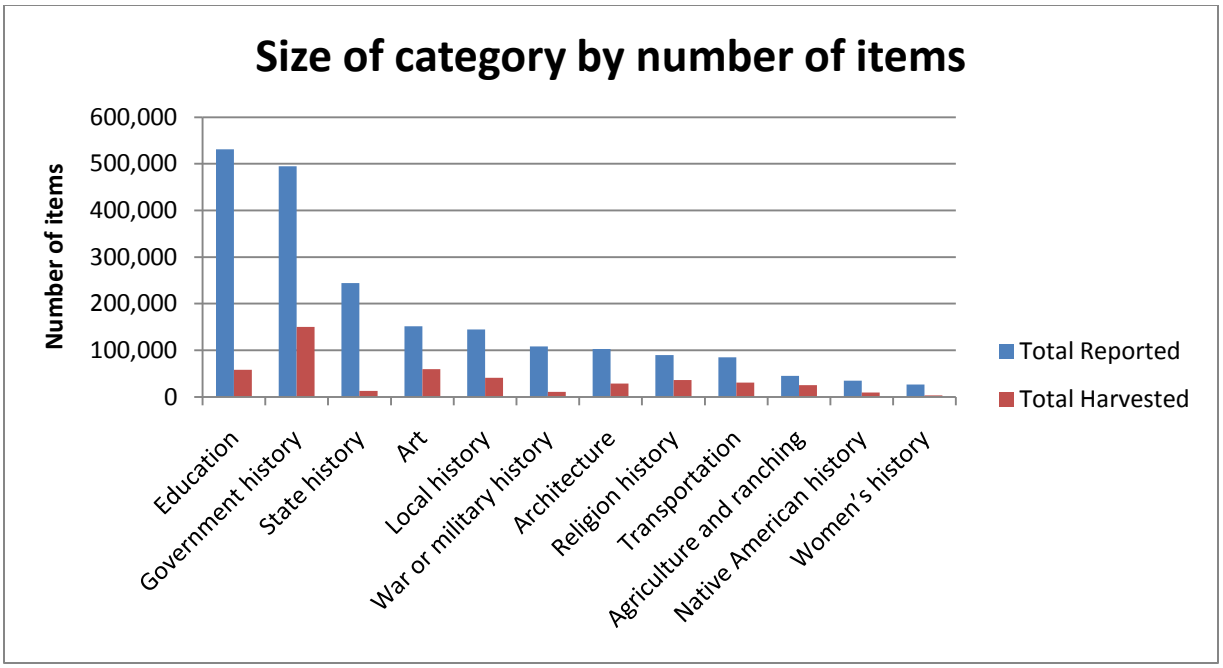
Because of differences in the nature of the categories and how they are assigned (in addition to variations in subject description practice), this ranking is an imperfect representation of the top subject strengths of the aggregate. Certain categories are problematic because they are often assigned coincidentally with other categories or because some topical umbrellas are simply larger than others. So these category comparisons, as I said before, should be taken as estimates.

In particular, I have deemphasized the top 2 subject strengths in the chart because I do not trust their comparability to the other categories. *Local history* is applied to terms that pertain, to any degree, to the history of a town, city, county, or other inhabited place below the state level. While it is valuable to know that the aggregation brings together resources about many dispersed but potentially related local histories, thereby making local histories more visible within a larger historical context, this category designation does not point to a unified topical strength in the same way that, say, the category of *military history* does. The *Education* category, which is often assigned to terms that refer to a specific educational institution, is similarly dubious as a pointer to a unified subject. My intuition is that these two categories in particular encompass more topically diverse collections, or perhaps collections with less substantial topical relationships among themselves than those in other categories. I wonder how this might be measured or accounted for, and how we might come to a better understanding of category overlap and co-occurrence.

The following chart compares subject strengths identified in the current evaluation to the previous top subject strengths, from the April 2010 evaluation.

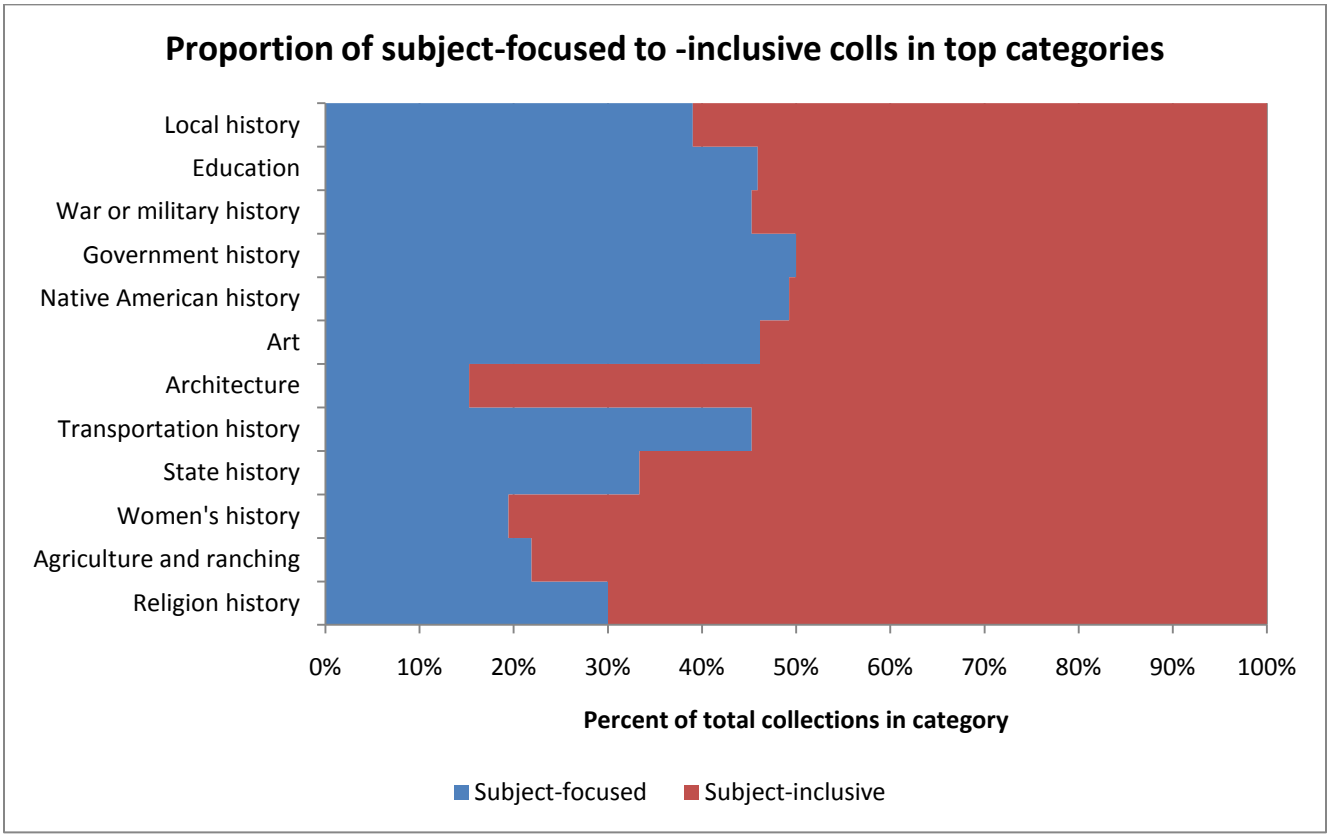


- As we can see, certain categories (*Native American history* in particular) appear to have shrunk with the change in evaluative method, while others have been newly identified or grown. This graph illustrates how the new evaluation method establishes a new baseline for assessing future growth in subject representation.



- In comparison to above graphs of subject strengths, this graph illustrates how measurement of subject strengths at the collection level does not yield a complete picture of the topicality of the aggregation, as many users will encounter it (through item-level search). This graph also illustrates the discrepancy between size of collections as they are described and size of collections as they are available for harvest.
- Most of this evaluation is done at the collection level. Viewing the top 12 strengths (where top 12 is based on collection counts) in terms of the number of items we harvest and make searchable reveals a striking difference in ordering of the top 12 subjects within the aggregation. However, these item counts as an indicator of subject strength must be mistrusted, because saying we have 500,000+ items in the category of *Education* actually means that we have that many items *in collections* that are categorized as *Education*. We do not yet evaluate subjects at the item level, due to the heterogeneity of item-level subject representation across the aggregation. This is something for SALT to explore.
- Discrepancy between reported and harvested sizes may suggest several things:
    - Collections may be incompletely digitized or shared, for any number of reasons. For example, digital collections may be provided to an aggregation as lures back to an institutional context (digital or physical) in which more is available for interested users.
    - Collection admins may report the size of a collection as larger than what is harvestable because (1) parts of the collection are not described in harvestable metadata records; or (2) what collection admins understand as a unit of the collection, as or as a resource gathered into the collection, may not align with units practical for sharing. The discrepancy between logical and the pragmatic, or something.
    - We need to implement alternative harvesting methods, since much of this difference may be attributable to lack of OAI-PMH provider implementation. Or we may simply need to catch up on our harvesting.

- The largest harvested category is Government, which is also the category with the highest proportion of subject-focused to subject-inclusive collections – see next chart. Intriguing. What does this convergence of factors mean? Is Government actually the strongest subject in the aggregation?



- This graph shows ratio of subject-focused to -inclusive collections among the top categories. Categories are ordered by strength, in terms of number of at-least-inclusive collections. In the top several categories, around 50% of collections in the category are subject-focused, the other 50% subject-inclusive.
- We hypothesized this kind of balance as a feature of budding thematic research collection, in our most recent DiCE paper. The subject-focused vs. –inclusive pattern becomes more sporadic as the size of category decreases. This is a possible place for prioritizing future collection development.

On average, each collection is assigned 1.39 categories. The most categories that any one collection has been assigned is10; two collections each have 10 categories:

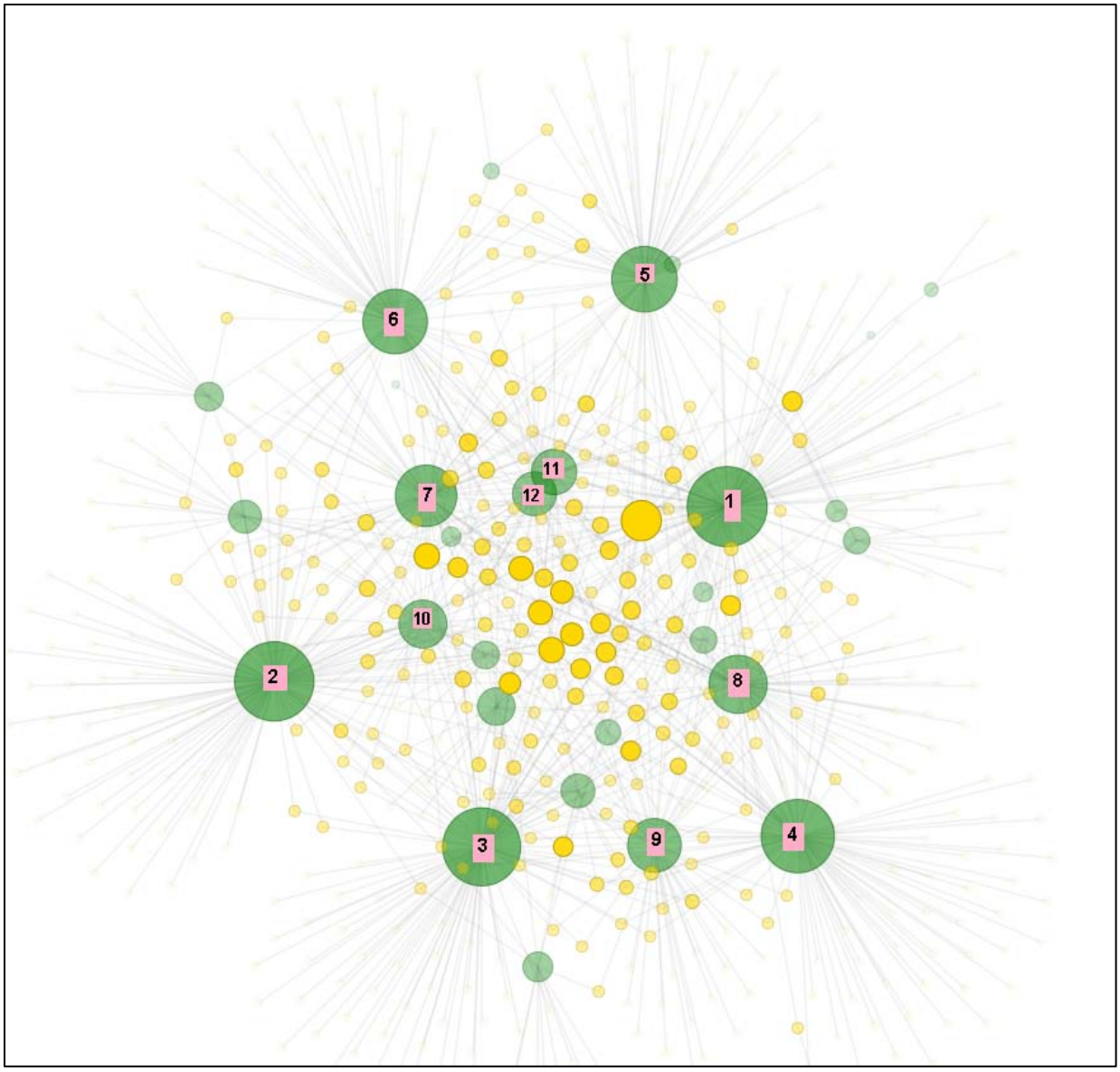
- "North Carolina Experience, Beginnings to 1940", a "Documenting the American South" collection from the University of North Carolina, Chapel Hill. This is a diverse collection intended to be comprehensive of NC state history.

- "Capital District Library Council Digital Collections" From the Capital District Library Council, which is an aggregation from libraries, archives, and museums in 10 New York state counties; not a huge collection, but broad subject coverage.

I have not fully understood the implications of differences between this method and Oksana's more detailed and manual evaluative method. The following are my current thoughts about why the data looks quite different, about what amounts to a compromise between precision and labor:

- Starting evaluation afresh as an term-by-term analysis of decontextualized terms means that some of our preexisting assumptions about DCC strengths were not imported into the current evaluation. This is why the 'top strengths' landscape looks so different. Analysis is, in some ways, more granular.
- Current method of term-by-term analysis should allow easier transfer of procedures and decisions between evaluators (easier to document method; reliance on standard queries; more flexible if we change categories; better at efficiently accommodating updates to old metadata, potentially)
- Changes how subject-focused vs -inclusive assessed (because how do you quantify the idea of 'focus'?)
- For now, term-by-term omits Description fields and item-level records (assessment of which entails unaffordable amounts of human labor, unless SALT has something to say about it)
- On the upside, term-by-term may be more successful at discovering patterns of subjects that may not appear to a human interpreter to be prominent in any one record, but which emerge as significant on the whole (e.g. *Local history*)
    - But term-by-term will miss subjects that are implicit, and may take terms out of context; Oksana's results will for the most part be more in line with our expectations

Category overlap within collections is an interesting place for future analysis. For starters, Peter visualized the collection/category relationships in the following network graph. While static here, the dynamic version of this graph is available and more interesting. It is imperfect but good for getting a quick read on the data. Each green node represents a category: its size indicates the number of distinct collections that link it. Yellow nodes are collections: their size indicates the number of distinct categories that have been assigned to them. Lines between nodes may be thought of as strings: strength of connection between two nodes pulls the nodes closer together. Categories that appear to cluster together usually have more shared collections between them (a.k.a., they co-occur commonly). Certain outlying categories with no connections to this graph are not shown. It is interesting that smaller categories are more central to the graph. This kind of visualization will be worth a more detailed analysis in future evaluations. See Katrina or Peter for a demonstration or copy of the dynamic visualization.



Legend

1 Local history
2 Education
3 War or military history
4 Government
5 Native American history
6 Art
7 Architecture
8 Transportation history
9 State history
10 Women's history
11 Agriculture and ranching
12 Religion history

2c. Topic-level views

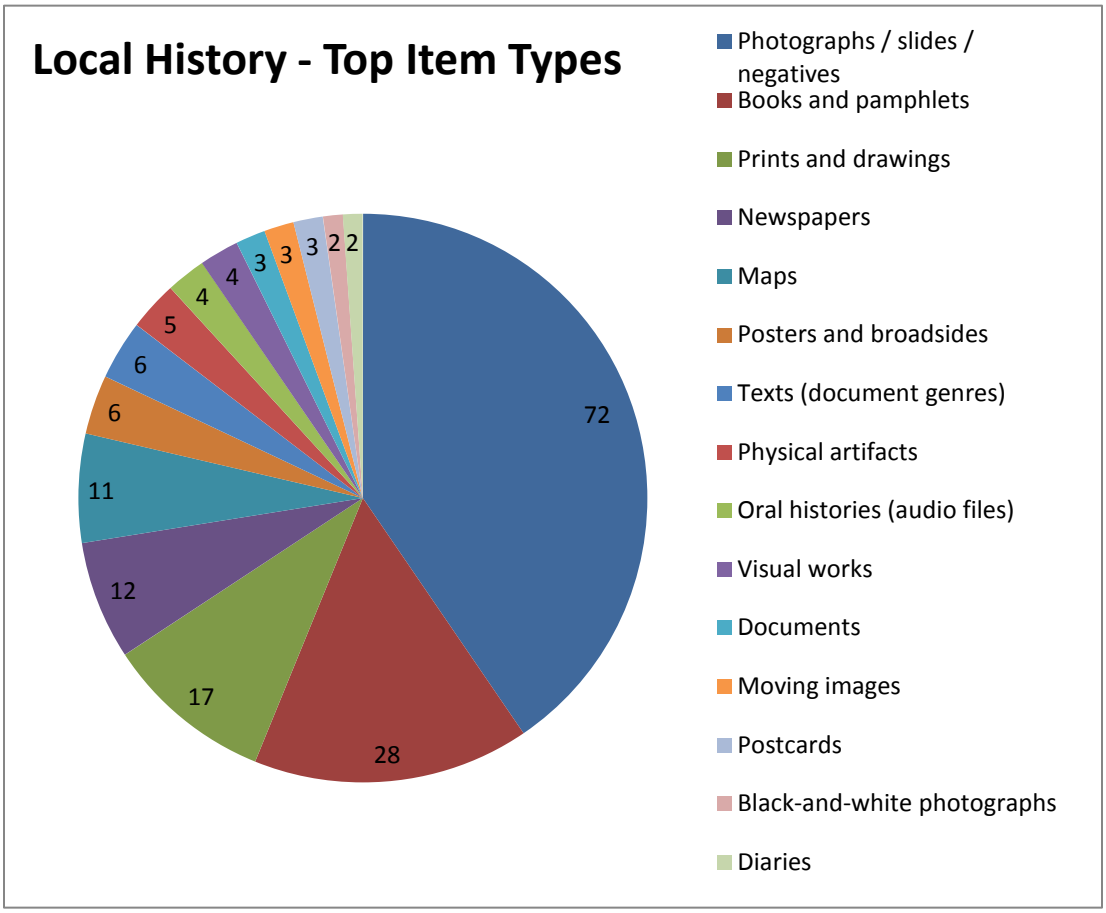
Local history

Overview

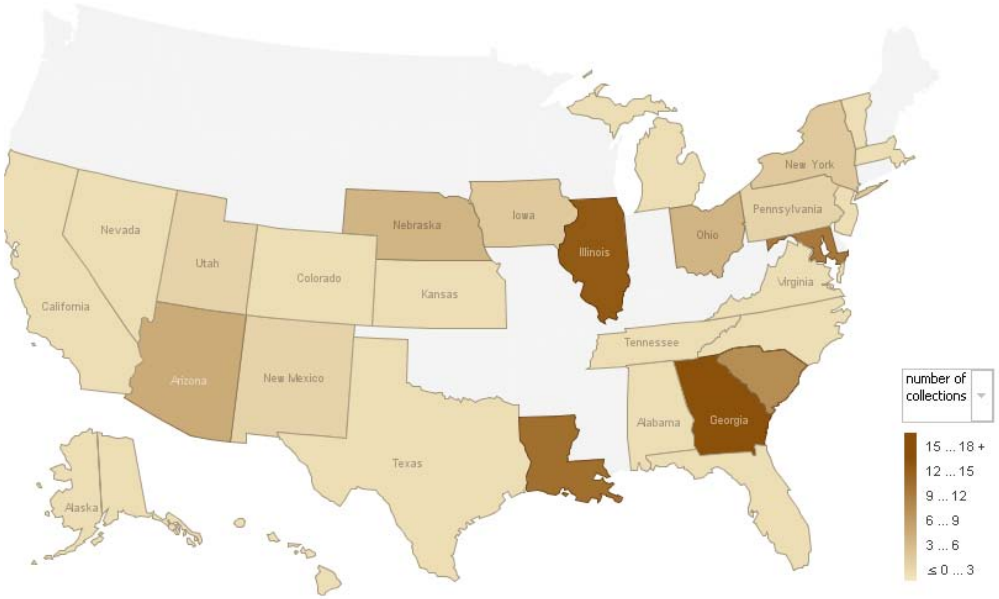
100 collections; 39 focused; 61 inclusive

Types

59 item types; 21 types in more than 1 coll



Geographic Coverage



- Coverage in 31 states
- City and county coverage within states would perhaps be a more interesting metric, but harder to get a clean read

Focused v inclusive

Many of the inclusive collections do focus primarily on the local history of an inhabited place; often they focus on a specific facet of local history (e.g. architecture, or the experience of a particular minority group) that happens to coincide with another category. Thus their intellectual focus may be on a single subject, but that subject falls between two categories in my scheme. So they're called inclusive. Similarly, a lot of the apparently focused collections would be inclusive according to some other subject organization. On the whole I think the distinction is useful but not perfect.

- Most subject-focused collections center on a town (historic maps, oral histories, pictures) or personal/family collection
- Most subject-inclusive collections are either broadly scoped image collections with pieces focused on localities, or else more locally focused collections that have a primary or secondary focus on another prominent topical category of the aggregation (mining, disasters, architecture, entertainment, religion, slavery)

Category co-occurrence

Co-occurrence: *Local history* as a category co-occurs with *Agriculture and ranching*, *Entertainment history*, *Government history*, *War or military history*, *African American history*, *architecture*, *disasters*, *women's history*, *art*, *religion history*, and *civil rights*. Unexpectedly, *Local history* demonstrates low co-occurrence with other categories often (relative to co-occurrence in other categories – see below); only about 15 indistinct instances of co-occurrence.

Subcategories

Local history does not yet have any subcategories.

Subject-specific strength factors

- This is a place for further analysis.
- Maybe:
  - Distinct inhabited places (cities/towns), or some other finer grain of coverage than state (e.g. county)
  - Dispersion of coverage within states

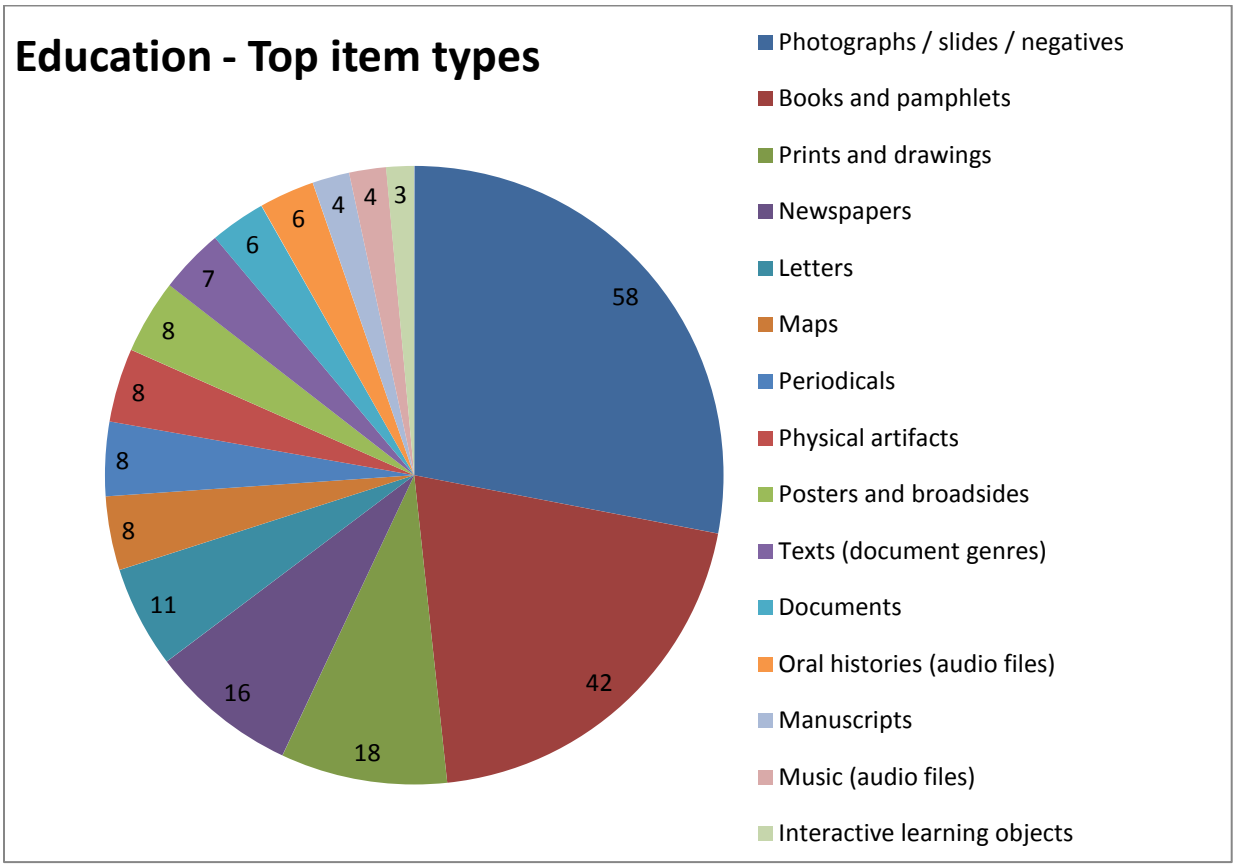
Education

Overview

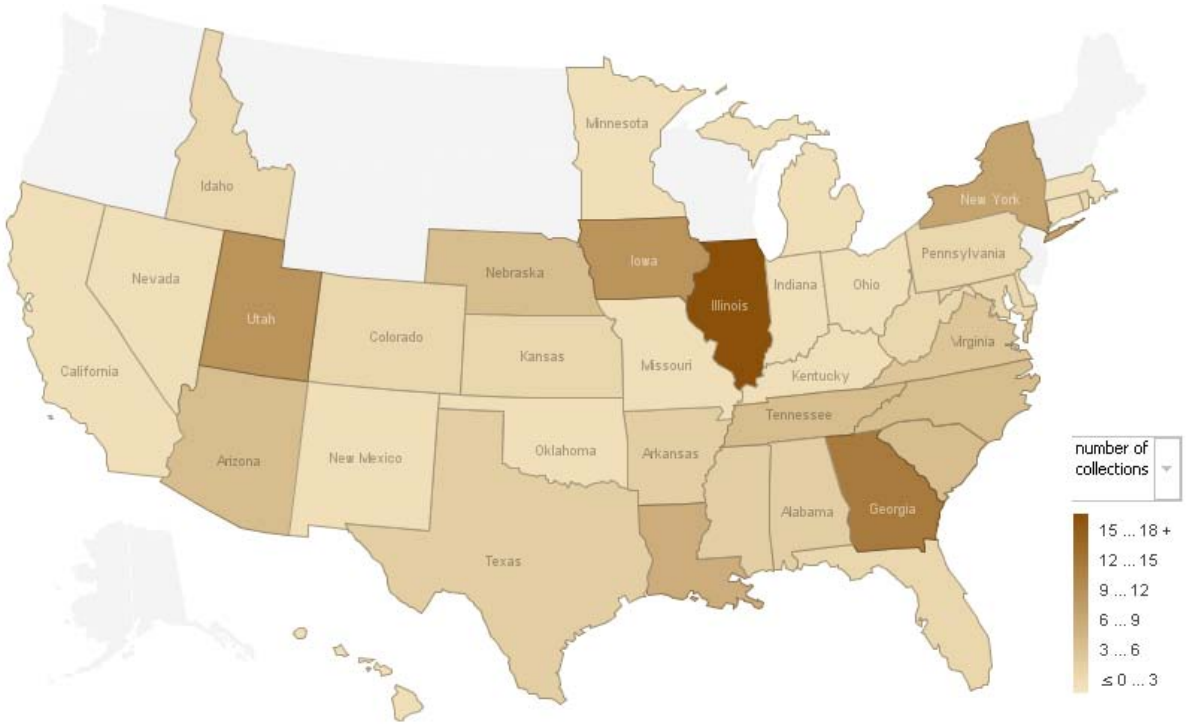
98 collections; 45 focused; 53 inclusive

Types

65 types; 22 types in more than 1 collection; 17 in more than 2 collections; more diverse than local hist, but not much. Appears to be a more even distribution between text and image than in *Local history*.



Geographic coverage



- Coverage of 38 states, pretty widespread

Focused v inclusive

- Focused collections mostly to come from what seem to be pointed acts of campus archive digitization (e.g. yearbook archives, news archives, or photo archives of a particular campus, alumni catalogs, etc.). One or two are more broadly scoped, e.g. “Education in Illinois: a collection of digitized books”.
- Inclusive collections appear to be regional or national photo collections, spanning many topics and regions, which include photos of educational institutions; full campus archives with a probability of encompassing more than one subject category; or education-focused collections that also deal heavily with another subject category (especially race or gender in education).

Category co-occurrence

Categories that co-occur with *Education* include *War or military history*, *African American history*, *Architecture* (usually pointing at resources related to school buildings), *Athletics*, *Native American history*, *Women’s history*, and *Art*. None of these is very dominant. In fact, relative to its size, it’s my impression that *Education* exhibits fairly low co-occurrence, with 30 total instances of co-occurrence.

Subcategories

In my subject organization, *Education* has 28 subcategories. These are topical areas in which items within a category pool. I repeat all the caveats I gave for categories at the outset of this document: lack of rigor,

formality, all bottom-up, lots of overlap, etc. In this case, these caveats don't matter too much since we're not doing any real quantitative analysis on subcategories. But it's interesting to note that "Universities and colleges" is by far the most common term on this list. That subcategory alone may account for *Education*'s high ranking in categories overall. That category is assigned to any term that pertains to a particular university or college; thus this subcategory has the features of the *Local history* category – not so much a unified topic as a shared tag. These are the 28 subcategories:

Academic libraries	Learning
Adult education	Lesson planning
ALA	National libraries
Coeducation	Libraries
Education commissions	Military education
Education history	School buildings
Education legislation	Schools
Educational associations	Segregation
Educators	Student activities
Folklore and education	Student publications
Fraternities	Students
Higher education	Universities and colleges
Integration	Art education
	Women education

None of these, except "Universities and colleges", as of yet appears to have sufficient size and diversity that it could be an emerging strength in its own right.

Subject-specific strength factors

- Discrete institutions?
- Education-related events?
- Or things like lesson plans?
- Hard to guess without having more user insight…

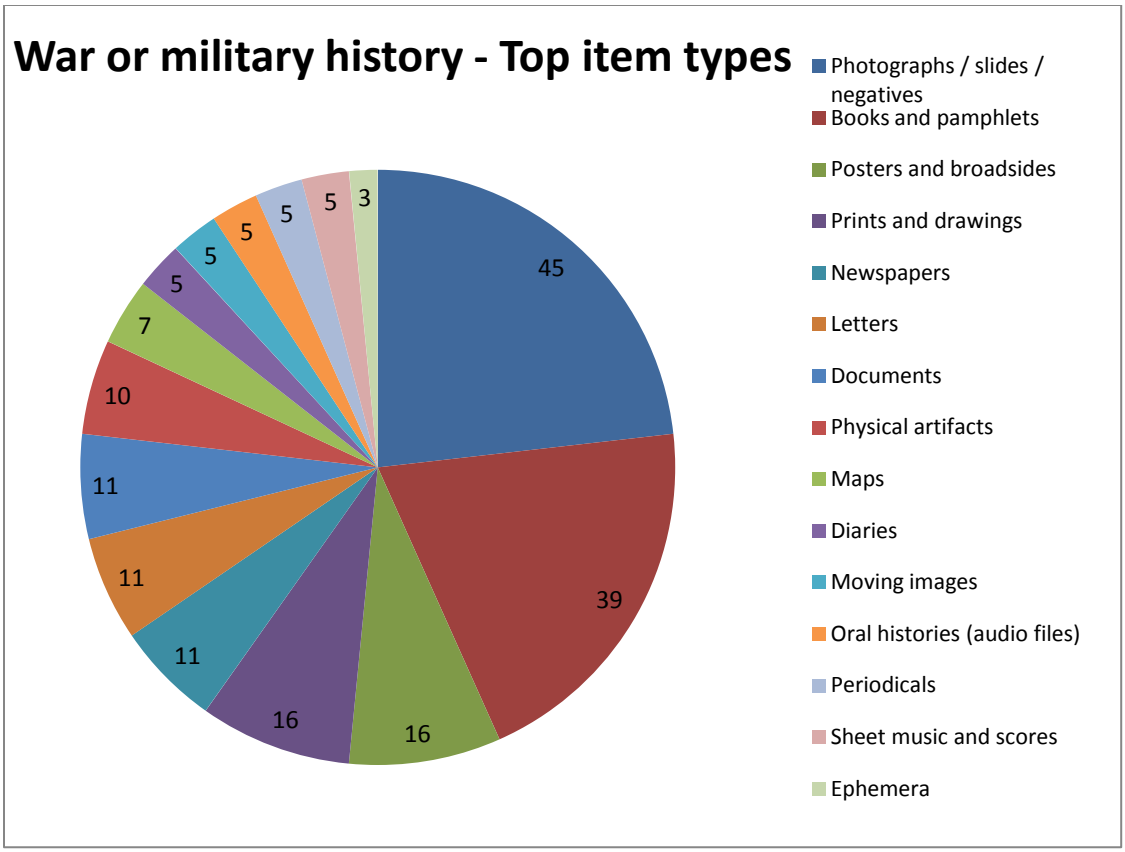
War or military history

Overview

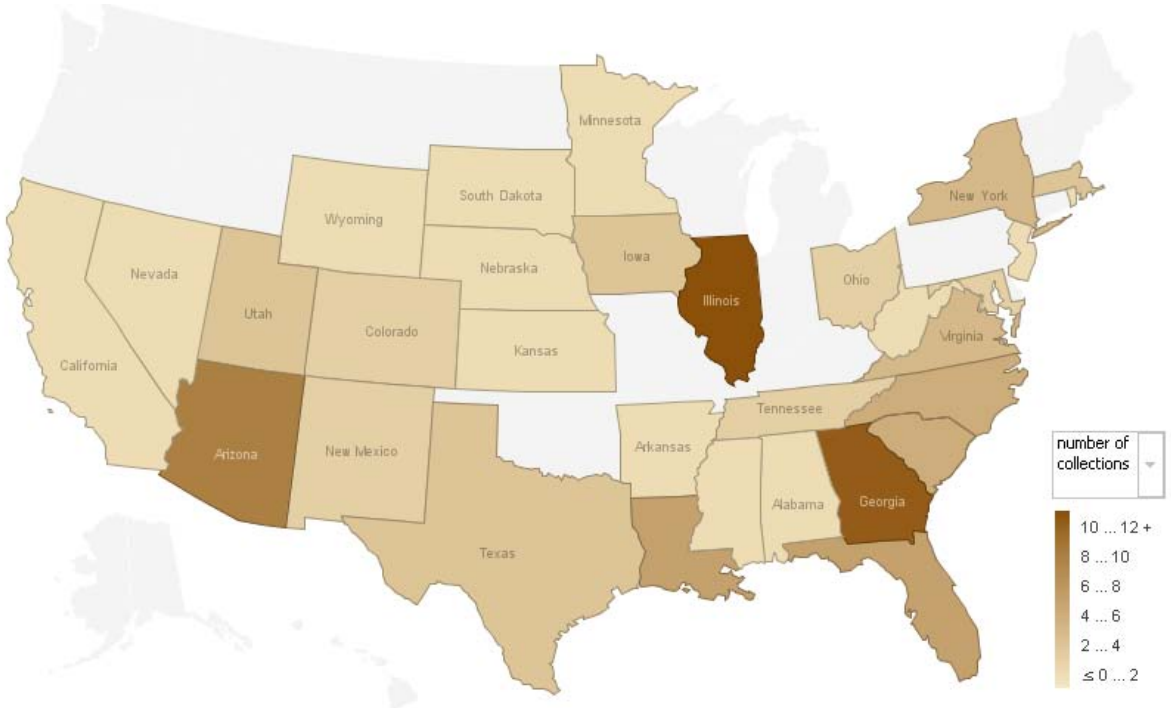
95 collections; 43 focused; 52 inclusive

Types

50 distinct types; 21 in 2 or more collections. Eclectic long-tail types include things like minutes, speeches, woodcuts, and World War I service cards. As in *Education* category, good balance of text and image, with photos being dominant as usual.



Geographic coverage



- Coverage of 31 states
- Stronger clusters in Illinois (focused on things like Lincoln and the Civil War, and local involvement with World Wars I and II), South (largely focused on the Civil War) and Southwest (largely focused on conflicts with Native American Indians, local efforts related to World War II, and governmental history of the state as related to military forces and events)

Focused v inclusive

- Focused collections on *War or military history* tend to be personal or family archives of people involved with or witnessing military history firsthand; otherwise these collections are targeted collections of well-defined scope, usually on military events or relevant locations (some on the scale of some facet of a war, others on a particular camp or region).
- Inclusive collections are very diverse. Some are particular-war-focused but draw in other categories. Some are focused on a particular historical figure (Lincoln, Tubman). Some are personal or family archives.

Category co-occurrence

- Occurs with 10 other categories. 23 total instances of co-occurrence. .
- Categories are *Asian American history*, *Education*, *Government history*, *Religion history*, *Transportation history*, *Local history*, *African American history*, *Architecture*, *Slavery* and *Women's history*. None of the co-occurring categories appears to be dominant.

Subcategories

- *War or military history* has 44 subcategories. Again, as I said above, there's plenty of overlap here and imperfect semantic comparability…

Cold War	Navy
international military	Persian Gulf War
Korean War	North Carolina
Marines	personnel
Mexican Revolution	prisoners of war
miliatry archeology	propaganda
military	regimental history
military camps	Revolutionary War
military courts	revolutions
military education	soldiers
military facilities	Spanish Civil War
military history	Spanish-American War
military industry	tribunal
military life	veterans
military museum	Vietnam War
military officers	war
military operations	war funds

military parks	War of 1812
military service	weapons
military structures	women in war
military transportation	World War I
national guard	World War II

Civil War is by far the dominant subcategory, represented in 32 of the 95 collections. In my estimate, at about 33% of the category, this of all subcategories (under any category) is the closest to being large enough that it merits category-like treatment (detailed evaluation, further division into subcategories). But I still think it's too small. 50% would probably be closer to category-worthy.

Subject- specific strength factors

- We had previously identified events as a likely subject-specific strength factor. As you can see, many of the subcategories are Event-centric.
- Beyond sheer number of events, more granular measures could be interesting: e.g. measures of coverage completeness and type diversity within a large-scale Event.

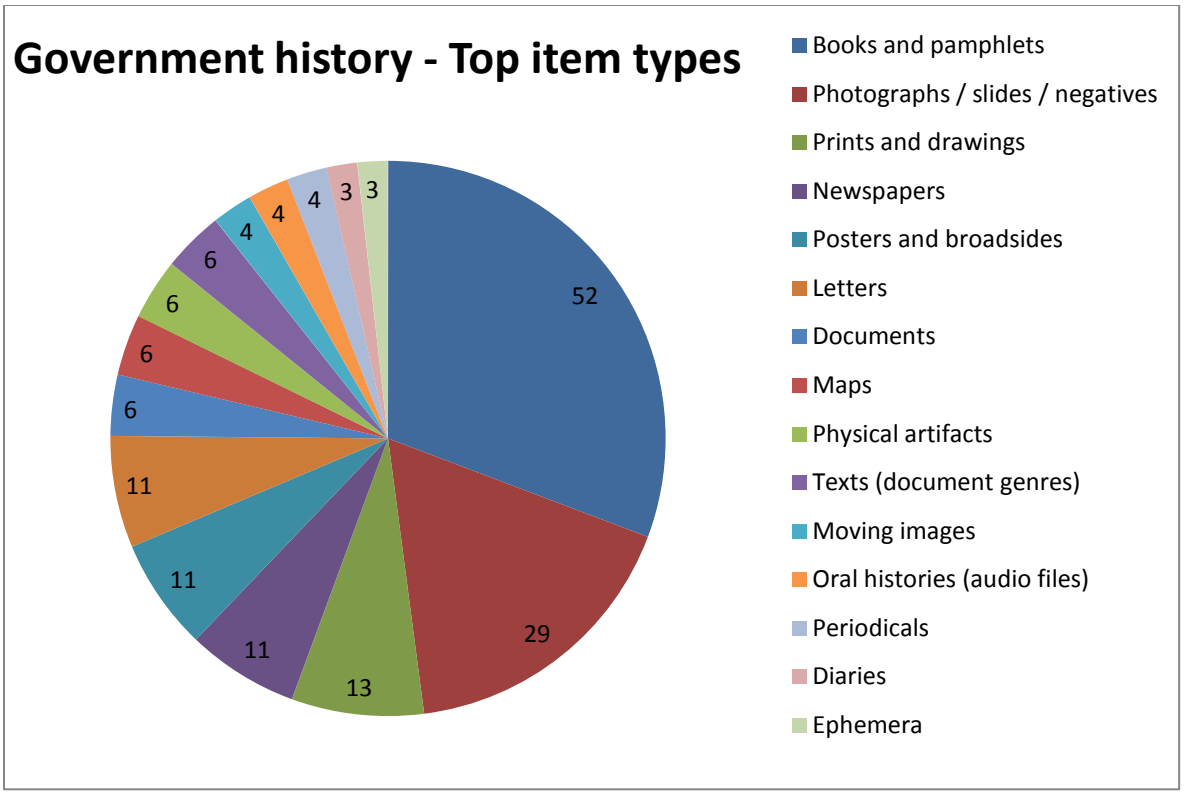
Government history

Overview

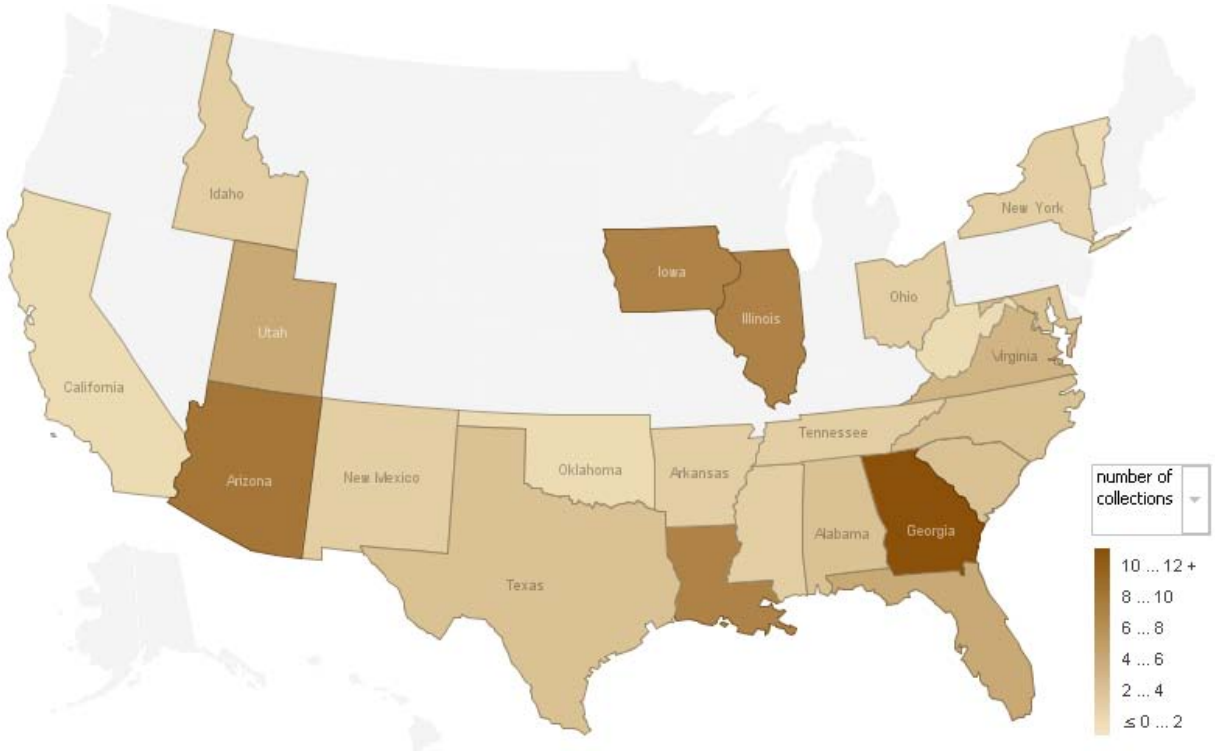
84 collections; 42 focused; 42 inclusive

Types

58 distinct item types; 22 appear in 2 or more collections. This is the only category I've assessed so far that includes more book collections than photo collections – the only category that appears to be predominately textual, if measured by number of collections only.



Geographic coverage



- Coverage of 24 states. Why the southern dominance? There is at least some correspondence here to state libraries we have worked with (which include but aren't limited to South Carolina, Texas, Arizona, Illinois, and Louisiana).

Focused v inclusive

- Equal parts focused and inclusive
- Focused collections tend to be select archives of government documents, often at a state level (a collection of historical constitutions from a state, a collection of registers or blue books, digitized government publications). Focused collections also include personal archives, eg correspondence collections, of politicians. Some are more purposive, theme-focused (rather than a digitized series or personal archive), such as the Civilian Conservation Corps Collection. It should be noted that many of the older collections in DCC/OH do not fully representative subject headings, so certain collections that appear as focused in this evaluation might not, if more care were taken with subject description. This is a place for improvement in the next eval.
- Inclusive collections tend to be broader government digitization projects (or at least more carefully described digitization projects) or subject-focused collections (often focused on historical figures, such as Lincoln) that happen to overlap with another category in this scheme.

Category co-occurrence

- *Government history* only co-occurs with four other categories, which are *Local history*, *War or military history*, *African American history*, and *Slavery* (note the overlap between the last two categories; this is a place for assessment in the next eval). There are only 12 instances of overlap. Very little compared to in other categories (not surprising, since focused/inclusive ratio is high, and because of the nature of subject description applied to government document collections in the aggregation).

Subcategories. The Government history category has 25 subcategories:

city planning or civil engineering	local government
commissions	national politics
constitution	national security
currency	political party
economics	politicians
elections	politics
federal programs	president
foreign policy	protests
government	revolutions
government officials	state government
government publications	state government publications
international politics	state programs
law	

Of these, state history and presidential history are the most prominent subcategories.

Subject-specific strength factors

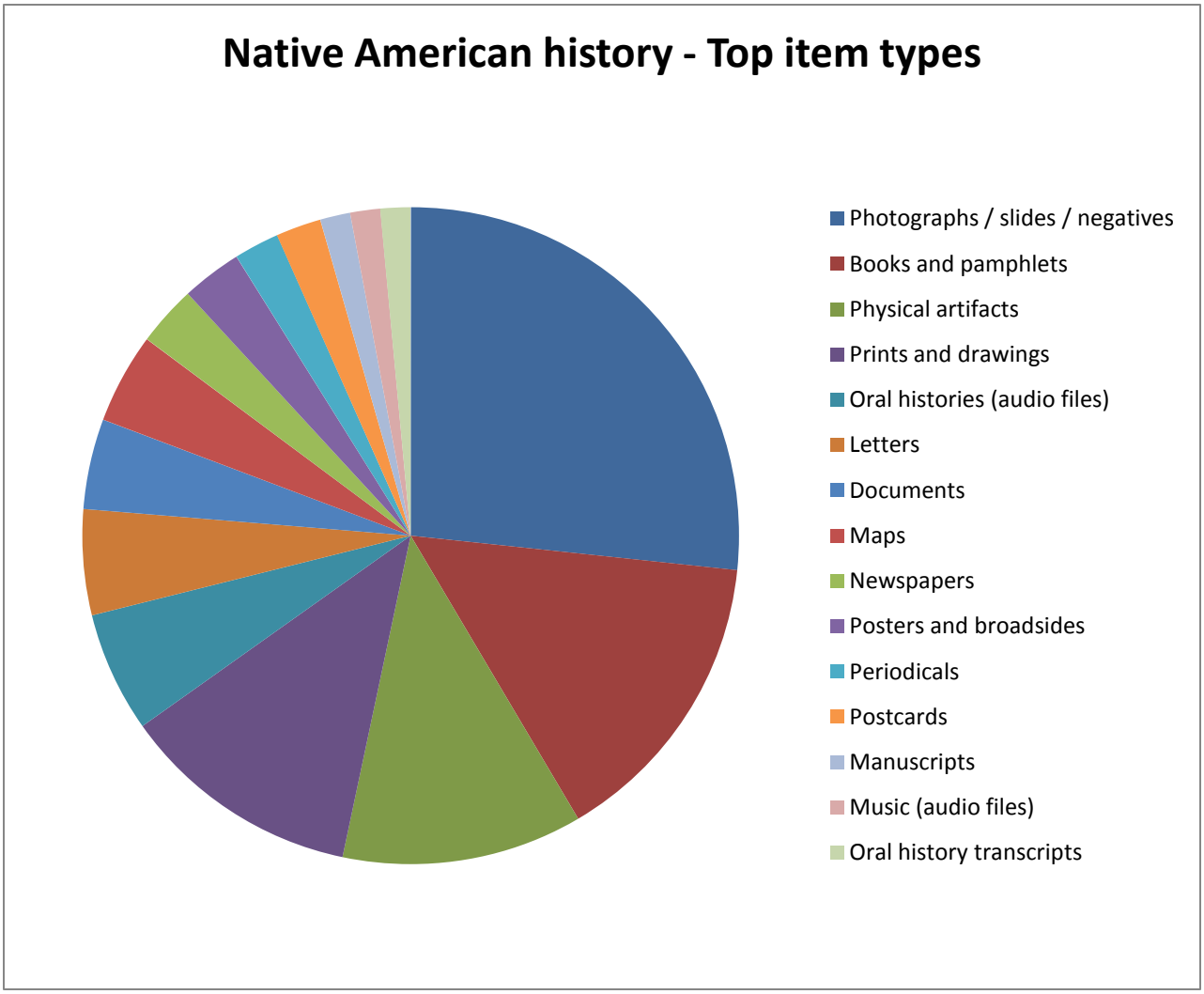
- Need brainstorming

Native American history

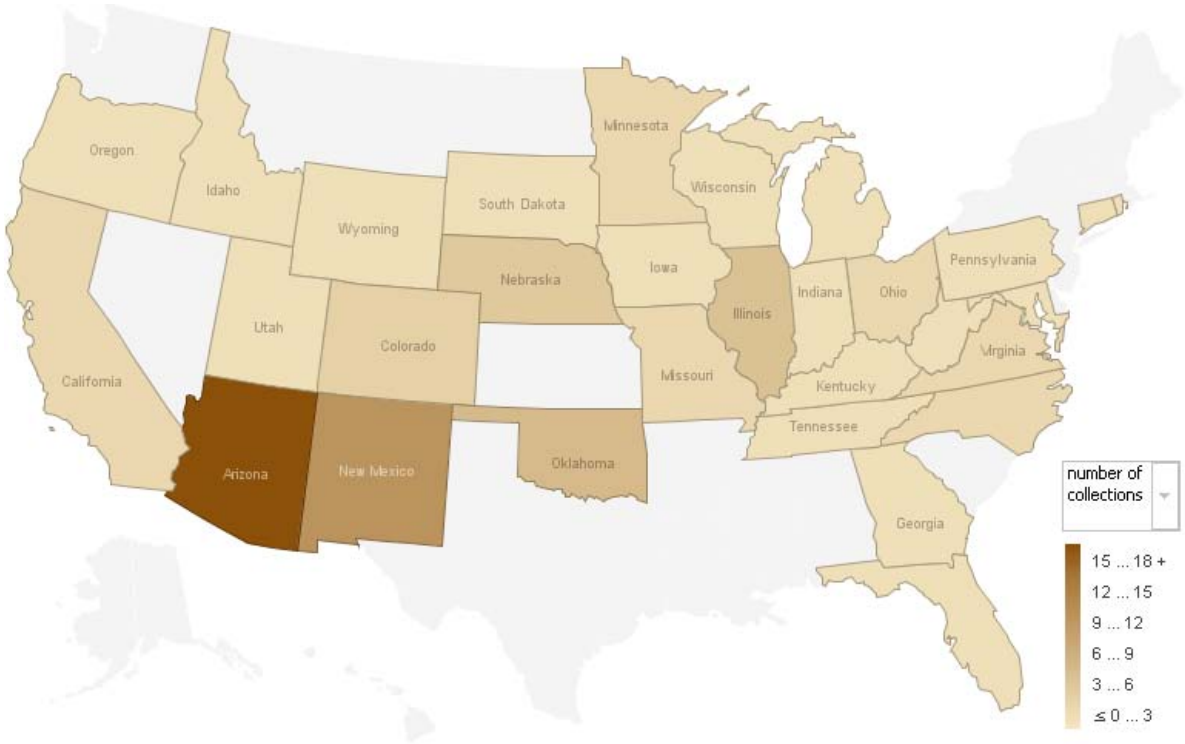
Overview

67 collections; 33 focused; 34 inclusive

Types



Geographic coverage



- 30 states
- Concentrated in Arizona (because of massive state aggregation from there)

Focused v inclusive

- Focused collections primarily come from academic and research libraries and centers. There are a few from museums. They focus on specific aspects of Native American history, art collections, and collections that revolve around historical figures/families.
- Inclusive collections are either more broadly scoped, focusing on regional histories that include Native American history, or else they are just as focused as the focused collections but happen to overlap with other categories in this local subject scheme.

Category co-occurrence

- Like *Government history*, *Native American history* has a low co-occurrence rate. It occurs with four categories in 12 instances: *Education*, architecture, civil rights, and art.

Subcategories

There are no subcategories yet. Depending on growth, and depending on how we want to assess subject-specific strength factors like, in this case, Tribes represented, this is a place for improvement in the next evaluation.

Subject-specific strength factors

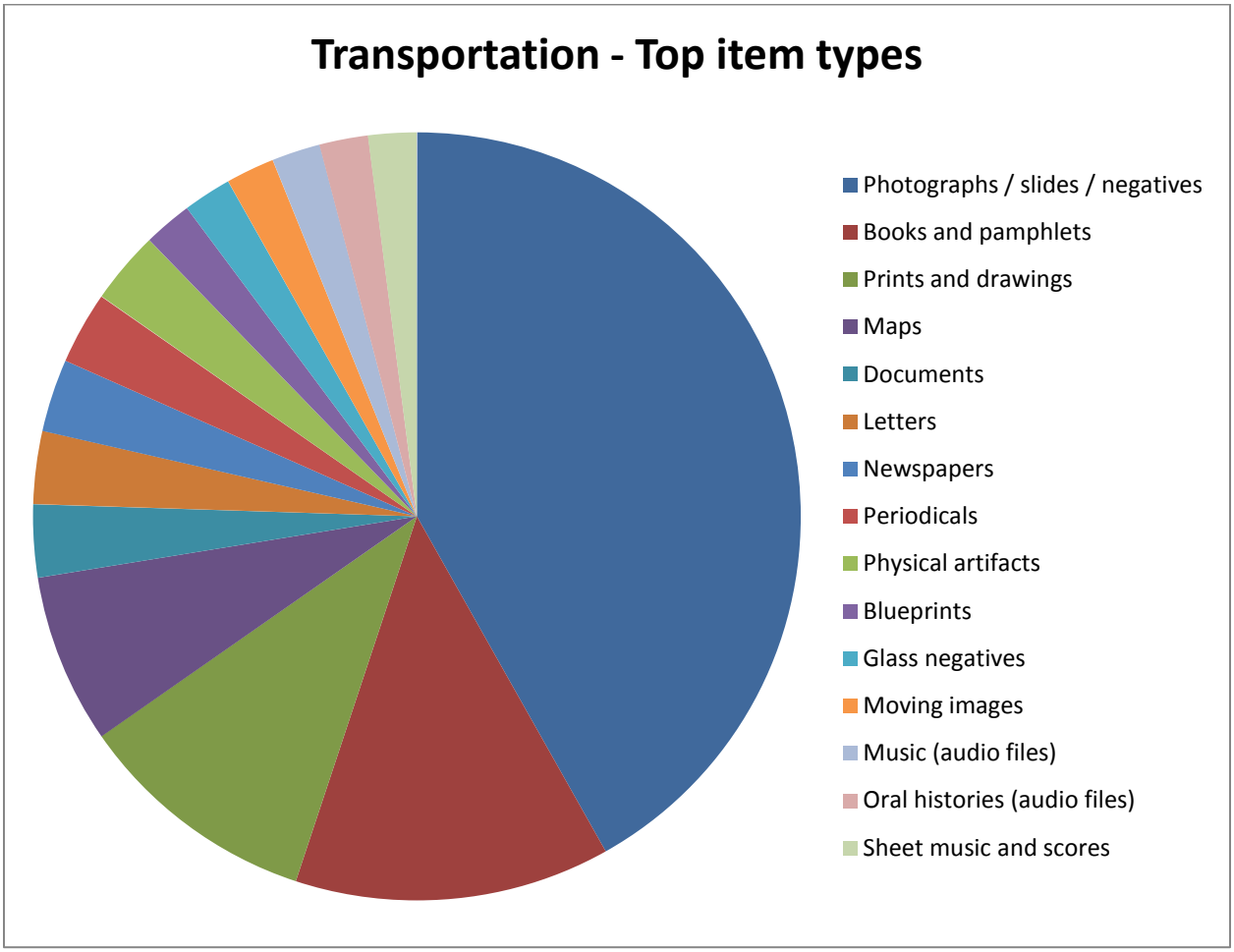
- Tribes. Oksana found 65 tribes represented. At the collection level, subject headings only describe 35. Attention to Description fields and item-level records will, apparently, reveal significantly more. This is a place for subject description improvement! (And something SALT group could help us with…)

Transportation history

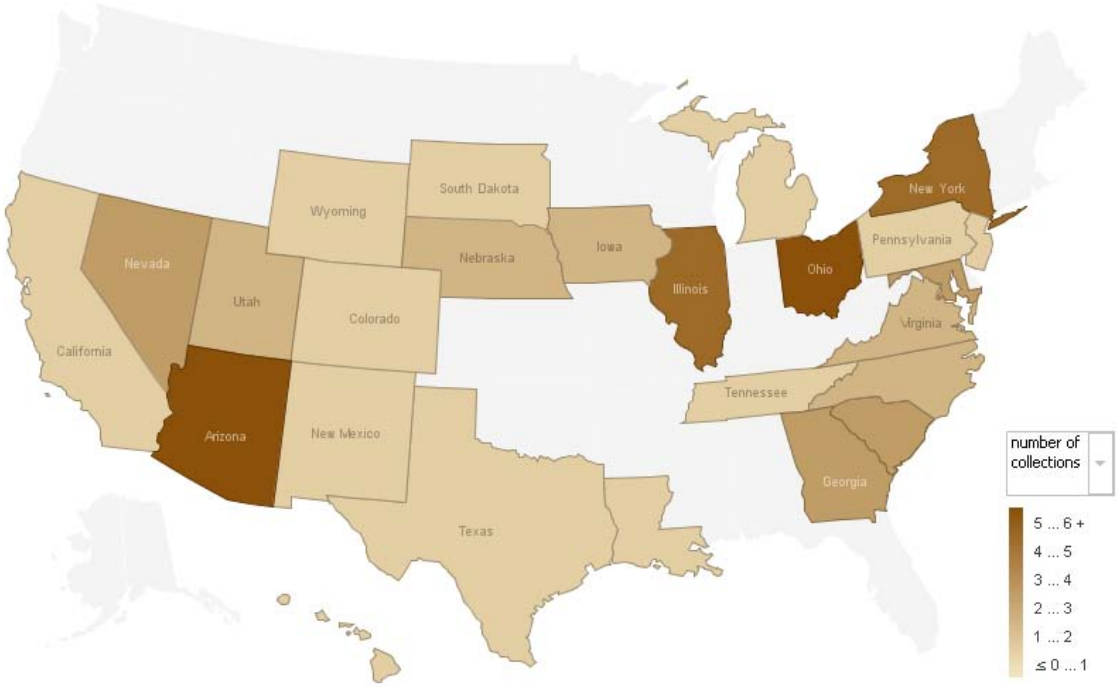
Overview

53 collections; 24 focused; 29 inclusive

Types



Geographic coverage



- 25 states
- Unusual pattern of transportation

Focused v inclusive

- Focused collections appear to be a mix of highly specific collections and more broadly scoped collections. Highly specific collections are focused on a particular make of some type of transportation, e.g. Pullman cars, or on some location, e.g. O'Hare or a highway. More broadly scoped, focused collections focus on things like the history of railroads in America.
- Inclusive collections range in the same way from specific to broad: from the Higgins Motor Torpedo Boat Diagram Collection to nationally scoped photograph collections (e.g. Cushman).

Category co-occurrence

- 6 instances of co-occurrence with 4 categories (low co-occurrence, perhaps, although we would have to take into account the smaller population of the category): *War or military history*, *architecture*, *oil*, and *art*. Pretty strange mix of co-occurring categories.

Subcategories

Transportation history has 22 subcategories, almost all specific to transportation types. This has potential relevance for determining subject-specific strength factors:

airplanes	ferries
aviation	public transportation

boats	railroad company
bridges	railroads
buses	roads
cars	shipping
electric railroads	ships
gliders	streetcars
military transportation	transit company
planning	transportation

Subject-specific strength factors

- Need brainstorming here. Diversity of transportation modes covered?

---

3. Next steps

- For future analysis: A size or other comparison of focused and inclusive in this case could be interesting. I have a guess that this category occurs commonly among collections that are part of supercollections (some level of aggregation below DCC level).
    - Also, disparity between subject-focused and –inclusive numbers in this eval and the last eval: what is causing the difference? Is there a weakness in presumptions about focus in this approach?
- Look further into implications of co-occurrence (and tangency). This is at the term level, right? What does that tell us, if anything? There must be a better measure of co-occurrence.
- Is there significance to why coverage appears the way it does within topics?
- Work on subject -specific factors
- Better analysis of uncategorized terms: any room for new category assignments? Exploration and travel history might arise from these, with a closer look.
- More on identifying contextual mass: small/large collection complementarity, density, diversity and interconnectedness
- Clean up terms (some mistakes, some places where minimal normalization (e.g. of punctuation) could make a lot of difference.
- Better normalization and then analysis of item type values.
    - Where item types appear in order for each category, relative to mean of aggregate?
- For future analysis: category overlap / co-appearance
- Return to question of termless and category-less collections. What other terms in use? What do they tell us?