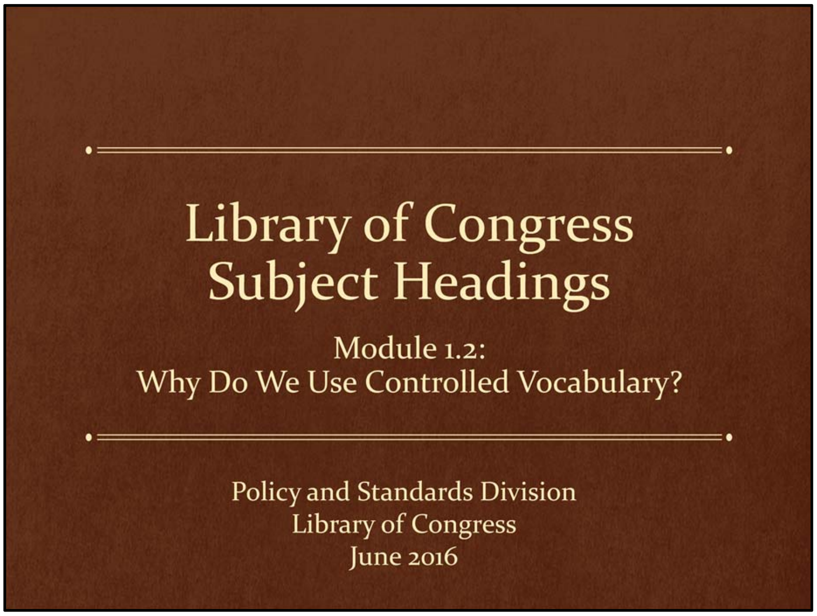
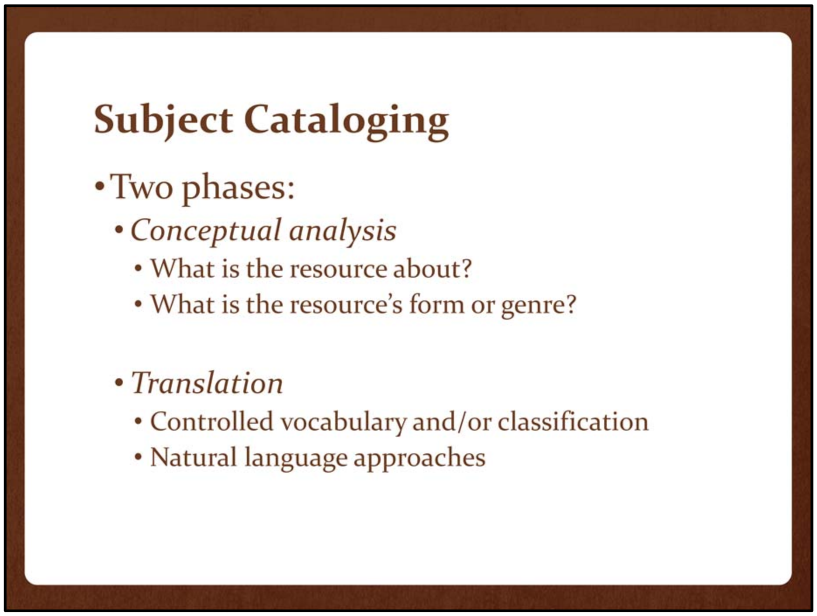


Library of Congress Subject Headings

Module 1.2:
Why Do We Use Controlled Vocabulary?

Policy and Standards Division
Library of Congress
June 2016



Subject Cataloging

- Two phases:
  - *Conceptual analysis*
    - What is the resource about?
    - What is the resource's form or genre?
  - *Translation*
    - Controlled vocabulary and/or classification
    - Natural language approaches

In this module we discuss the reasons why controlled vocabularies are used in subject cataloging. As we discussed in the previous module, the subject cataloging process involves two major steps.

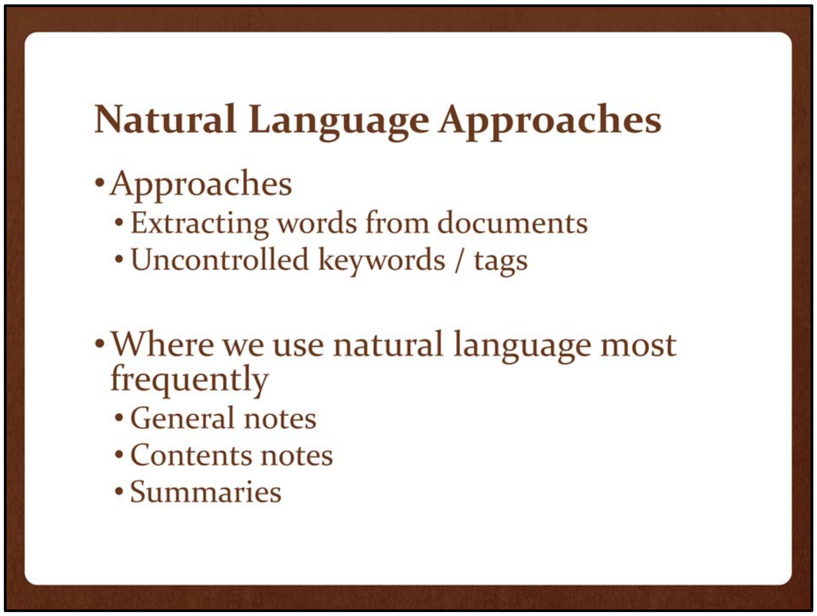
The first step is the *conceptual analysis*. This is the process of determining what a resource is about and what it is. In other words, catalogers must analyze a resource's content to identify and describe the subject matter, as well as to determine whether there are any significant, identifiable genre/form characteristics to be brought out in the analysis. This process will be discussed in more detail in the next module.

The second step is the *translation* phase. Once a cataloger has developed an understanding of the resource's aboutness and other characteristics, that understanding is translated into some form of description.

That description, however, *could* go in two different directions.

It is possible to describe the subject of a resource using either *natural language* or *controlled vocabulary*.

During this module, we'll talk a bit about each of these paths.



Natural Language Approaches

- Approaches
  - Extracting words from documents
  - Uncontrolled keywords / tags
- Where we use natural language most frequently
  - General notes
  - Contents notes
  - Summaries

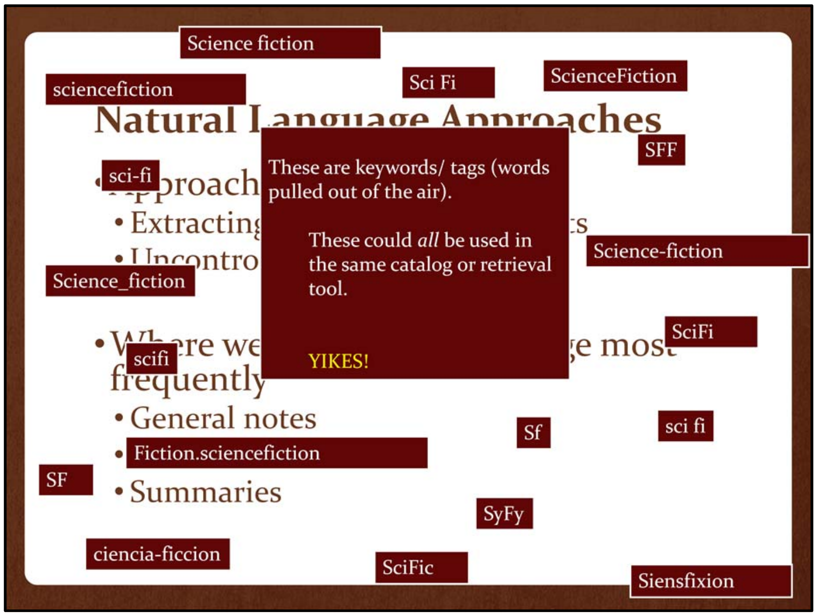
There are many ways that catalogers could describe the subject matter of documents.

They could pull words from the documents themselves, or catalogers could simply pull words out of their heads. We are aware, however, that these approaches lead to inconsistency.

People can use very different words to express the same idea. For example, just in English, sweet carbonated beverages, such as Coca-Cola or Pepsi, are referred to by various terms in the United States alone. In various parts of the US, that beverage is referred to as a soda, a pop, a soft drink, a soda pop, a tonic, or even as a “coke” (used generically, not referring to a specific brand).

And we *also* know that the same word (for example, bridge) can be used for two or more different concepts. The words chosen by a creator to represent a concept may vary throughout a text (using first one word and then a synonym, then another variation, and so on).

With this kind of inconsistency, it’s easy to see why extraction or keywords (otherwise known as tags) can be problematic.



Science fiction

sciencefiction

Sci Fi

ScienceFiction

Natural Language Approaches

SFF

- sci-fi Approach
  - Extracting
  - Uncontro

These are keywords/ tags (words pulled out of the air).

These could *all* be used in the same catalog or retrieval tool.

YIKES!

Science-fiction

Science_fiction

SciFi

- Where we scifi e most frequently
  - General notes
  - Fiction.sciencefiction
  - Summaries

Sf

sci fi

SF

SyFy

ciencia-ficcion

SciFic

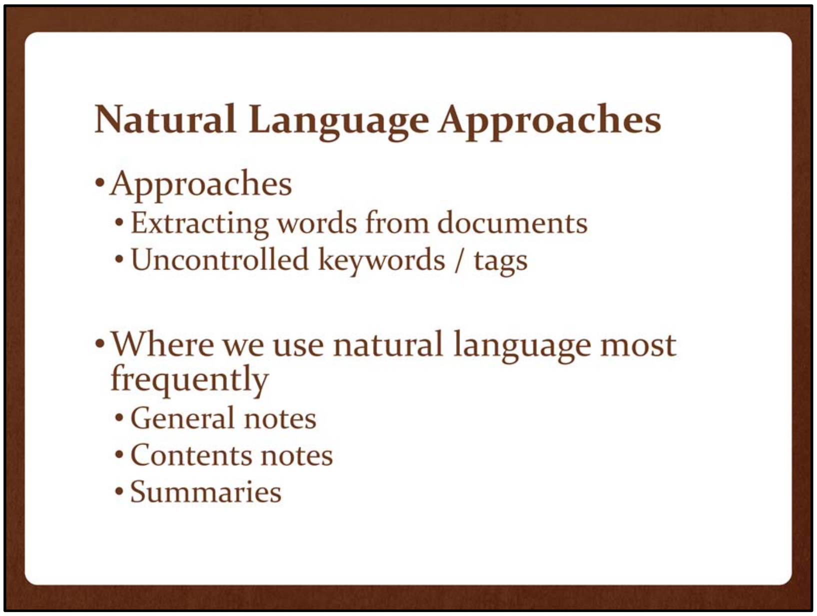
Siensfixion

In recent years, user tagging has become popular throughout social media but it also has been used for some specific library projects. For example, at the Library of Congress, the Prints and Photographs Division has invited the public to supply tags for digital historical images that it has uploaded to a FLICKR website. For a project of this nature, tagging has been helpful in identifying unknown places, objects, and people.

But in other contexts, tagging or general keyword assignment can be chaotic and unhelpful.

In this slide you can see just some of the many tags users have come up with to describe a particular genre of literature.

If this were the only approach to describing resources, it would separate the same type of resource into 18 different categories. That violates one of the goals of organizing information – to collocate (or bring together) like resources.



Natural Language Approaches

- Approaches
  - Extracting words from documents
  - Uncontrolled keywords / tags
- Where we use natural language most frequently
  - General notes
  - Contents notes
  - Summaries

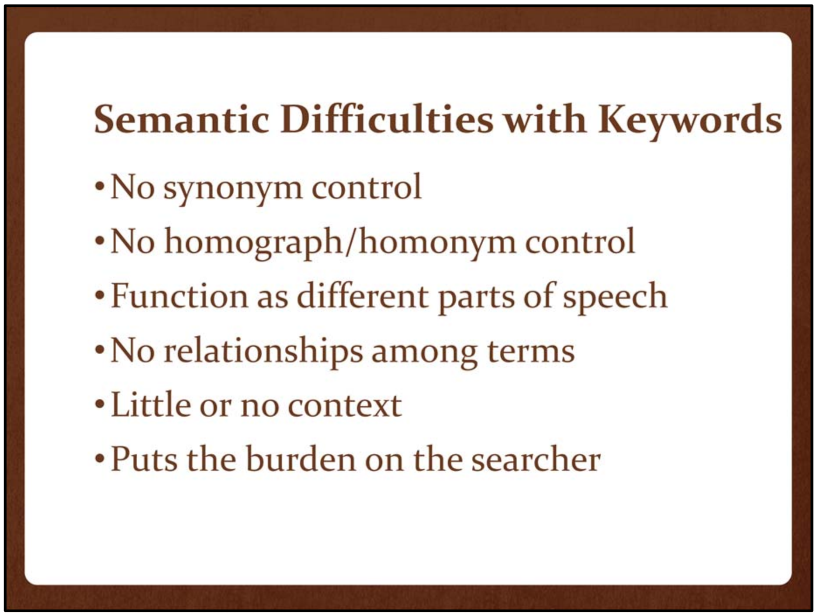
We do, however, use natural language in creating some library metadata.

We use it for all sorts of notes, particularly in contents notes where we transcribe the table of contents into our bibliographic description. We record the author’s words exactly in these cases; for example, we would never replace the phrase “soda pop” in a chapter title with the preferred term “soft drinks”!

We also see natural language used in summaries. If a summary is found in a publisher’s blurb on the back of a book, we will record it as it is written.

In some cases, the natural language terms found in summaries, abstracts, tables of contents, and other parts of the record can be useful in retrieving the resource if the terms used are different from the terms found in the controlled vocabulary.

In some cases, the searcher may choose a term that is widely used among metadata records and as a result get an overwhelming number of search results filled with many, many false drops. This certainly can impede efficient and effective retrieval.



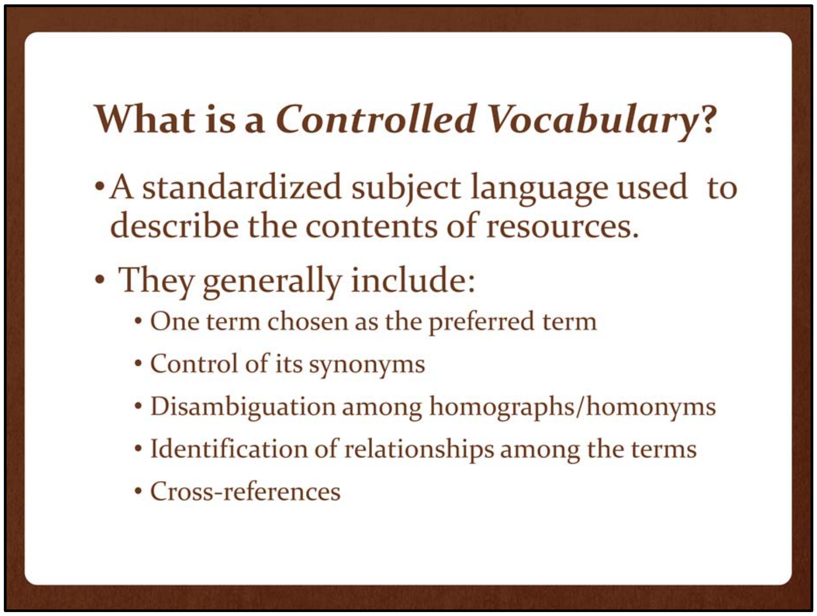
Semantic Difficulties with Keywords

- No synonym control
- No homograph/homonym control
- Function as different parts of speech
- No relationships among terms
- Little or no context
- Puts the burden on the searcher

Lots of semantic difficulties can arise when searching by keyword only. For example:

- We love synonyms, which are different ways of expressing the same meaning; this affects the way resources are written, the way resources are described for organizing purposes, and also the way we search.
- Virtually every word in the English language has more than one meaning or sense, and many of those senses even have more than one nuance.
- Many words can also be used as various parts of speech, such as nouns, verbs, adjectives, and adverbs. And, most search systems cannot yet distinguish among different meanings or various parts of speech.
- There also aren't relationships among keywords. There is a relationship between the terms *toys* and *dolls,* but the individual keywords are not connected.
- When keywords are used to describe resources, often those keywords lack any context. So we are left asking, How do those individual keywords relate to each other?
- And finally, searchers must come up with every possible word that is used to describe the concept if they are interested in getting all the pertinent materials. This can often require in-depth knowledge of the field.

Because of these types of difficulties, librarians, archives, museums, and other information institutions tend to favor the use of controlled vocabularies.



What is a *Controlled Vocabulary*?

- A standardized subject language used to describe the contents of resources.
- They generally include:
  - One term chosen as the preferred term
  - Control of its synonyms
  - Disambiguation among homographs/homonyms
  - Identification of relationships among the terms
  - Cross-references

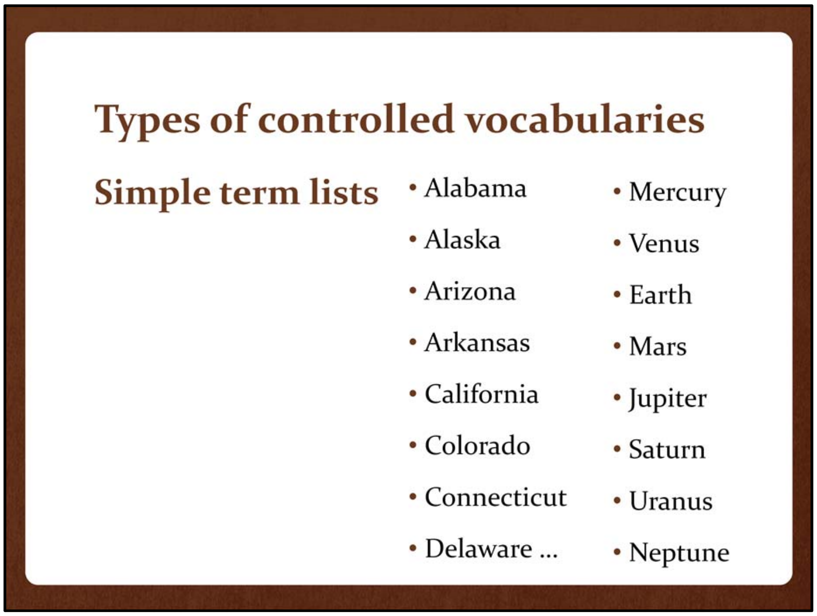
A controlled vocabulary is a list of authorized terms used to provide *consistency* and *uniqueness* among subjects in our descriptions of resources. The terms may be called subject headings, or descriptors, index terms, thesaurus terms, identifiers, or subjects.

Whatever you call them – all terms representing the same concept are brought together under one preferred term to provide collocation. For example, in LCSH we use **Young adults** even if the resource itself uses the phrases *Young people* and *Young persons*.

In LCSH, if a homonym is in popular use, we have to address it in some way, as well. For example, we use **Bridges** for the structures crossing rivers, but **Bridges (Dentistry)** for a partial denture. The term **Bridges** (unadorned) cannot be used for both.

Also, relationships among the terms are identified to create a syndetic structure (a network of relationships); we show that a doll is a type of toy and that the two terms representing those concepts have a hierarchical relationship.

Cross references are also created from unauthorized terms. The unauthorized terms point to the chosen (or preferred) term used to represent the concept. For example, we point from *Young people* to **Young adults** in LCSH.



Types of controlled vocabularies

Simple term lists

- Alabama
- Alaska
- Arizona
- Arkansas
- California
- Colorado
- Connecticut
- Delaware …

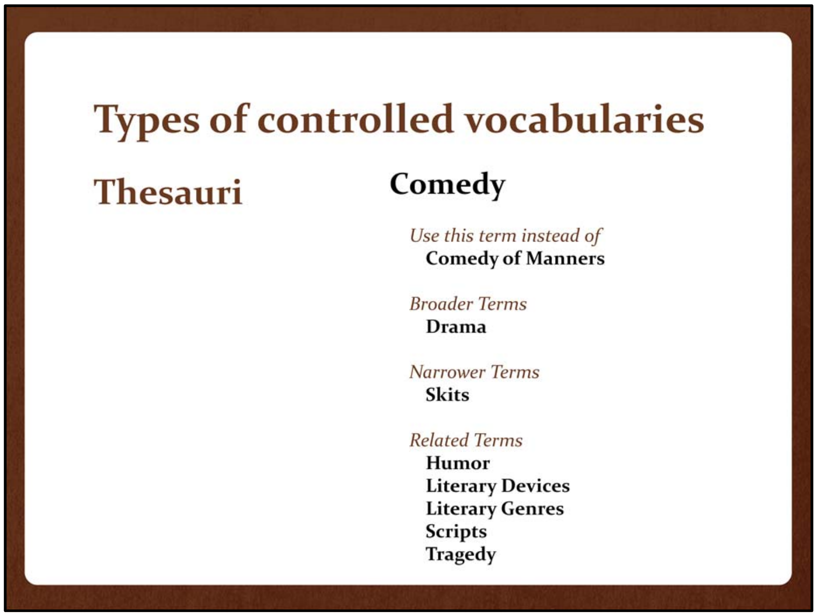
- Mercury
- Venus
- Earth
- Mars
- Jupiter
- Saturn
- Uranus
- Neptune

Controlled vocabularies appear in a variety of forms. One such form is the “simple term list” (sometimes called a pick list). This refers to a limited set of terms arranged as a simple alphabetical list or a list that is arranged in some other logically evident order.

These lists are *not* concerned with semantic relationships. They’re used to describe properties that tend to have a limited number of possibilities.

Examples might be geographic areas (maybe a list of countries or states or cities); maybe a list of languages; or perhaps a list of formats (which might include terms such as text or sound or image). They may be presented as pull-down menus in the cataloging system, so that they are available for easy use.

Two examples of simple term lists are seen on the slide. The first is an alphabetical list of states. The second is based on physical order or spatial contiguity. Simple term lists need not be hierarchical, if the list is short and there is some intuitive way of navigating the list, it can be useful without further structure.



Types of controlled vocabularies

Thesauri

Comedy

Use this term instead of
Comedy of Manners

Broader Terms
Drama

Narrower Terms
Skits

Related Terms
Humor
Literary Devices
Literary Genres
Scripts
Tragedy

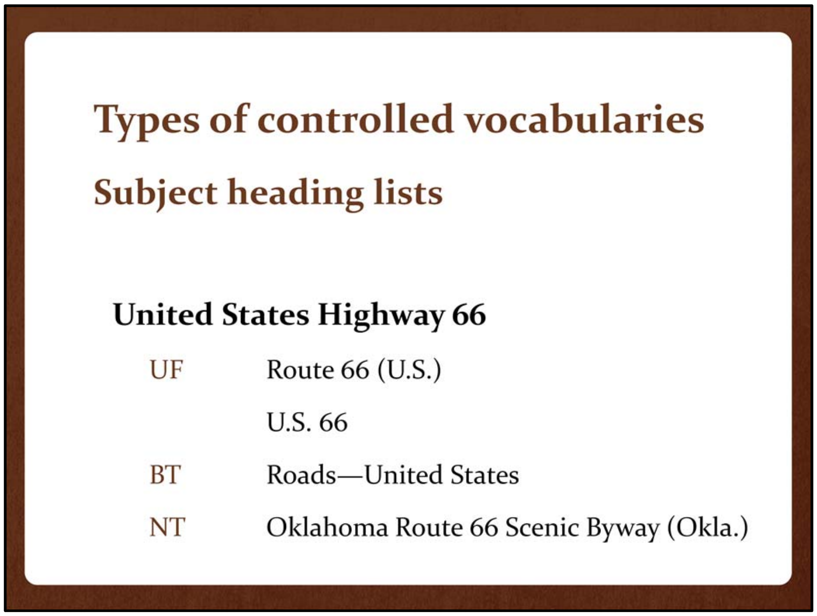
Another type of controlled vocabulary is the thesaurus.

Thesauri are lists of controlled terms that *do* include semantic relationships; the broader terms, narrower terms, related terms are identified within the structure of the list.

A thesaurus tends to focus more heavily on hierarchical relationships and it typically includes single terms or bound concepts, where maybe there is a phrase but it really represents a single idea.

A thesaurus tends to be a little narrower in scope than one of our next types of controlled vocabularies (subject heading lists, which we will talk about in a moment).

If you look at the screen you will see an entry from a thesaurus. This is from the ERIC Thesaurus, and it is the entry for **Comedy**. In this it identifies that this term is used instead of the term *Comedy of Manners*. It has a broader term of **Drama**, narrower term of **Skits**, and it also makes connection to some related terms: **Humor**, **Literary Devices**, **Literary Genres**, **Scripts**, and **Tragedy**. Generally you see mostly single-word terms throughout this entry, but you can see **Literary Devices**, **Literary Genres**, *Comedy of Manners* do include some phrases, but they are focused toward a single concept.



Types of controlled vocabularies

Subject heading lists

United States Highway 66

UF	Route 66 (U.S.)
	U.S. 66
BT	Roads—United States
NT	Oklahoma Route 66 Scenic Byway (Okla.)

Like thesauri, subject heading lists are lists of terms and multi-word phrases.

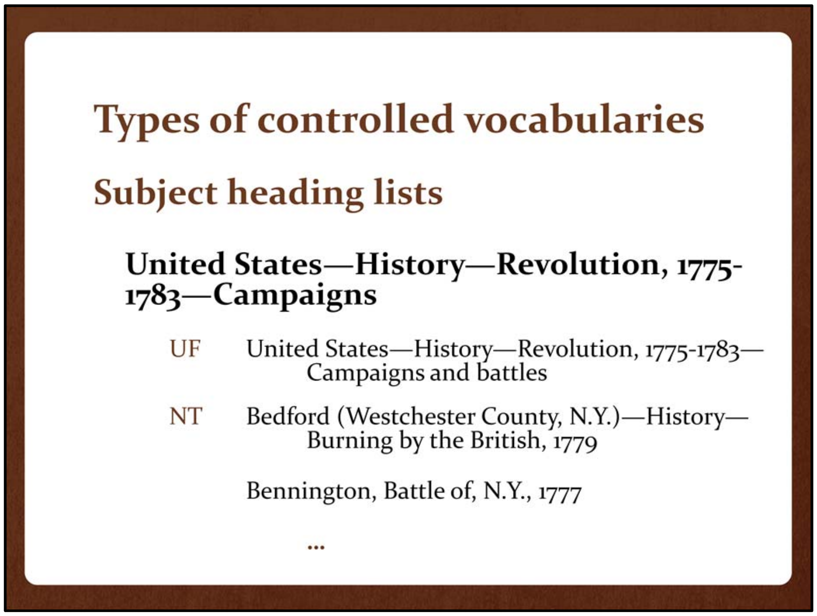
They also identify semantic relationships (such as broader term, narrower term, and related term).

They also generally allow for complex string construction, by allowing for subdivisions that create context for the main topic. Those subdivisions tend to represent additional topical elements, as well as geographic, chronological, and form aspects.

On the slide you can see a rather simple subject heading example from LCSH. The heading itself is **United States Highway 66**. Underneath that you will see two “see” references. They are marked with the initials UF, which generally stands for “used for”, and these are alternative ways of referring to the same entity. *Route 66 (U.S.)* and *U.S. 66* are synonyms or synonymous terms for **United States Highway 66**.

Also in the entry you will find a BT. BT stands for “broader term”. It is **Roads—United States**. What we have here is a broader concept. **Roads—United States** indicates that **United States Highway 66** is a road within the United States. Now you may think, “Oh, that’s really obvious!” Of course it is! However, this is a way of collocating entries throughout LCSH that are about roads within the United States.

There is also an NT on the screen. That NT stands for “Narrower term”, and that NT happens to be **Oklahoma Route 66 Scenic Byway (Okla.)**. This refers to a particular portion of **United States Highway 66**, and therefore has been designated a narrower term under the primary heading.



Types of controlled vocabularies

Subject heading lists

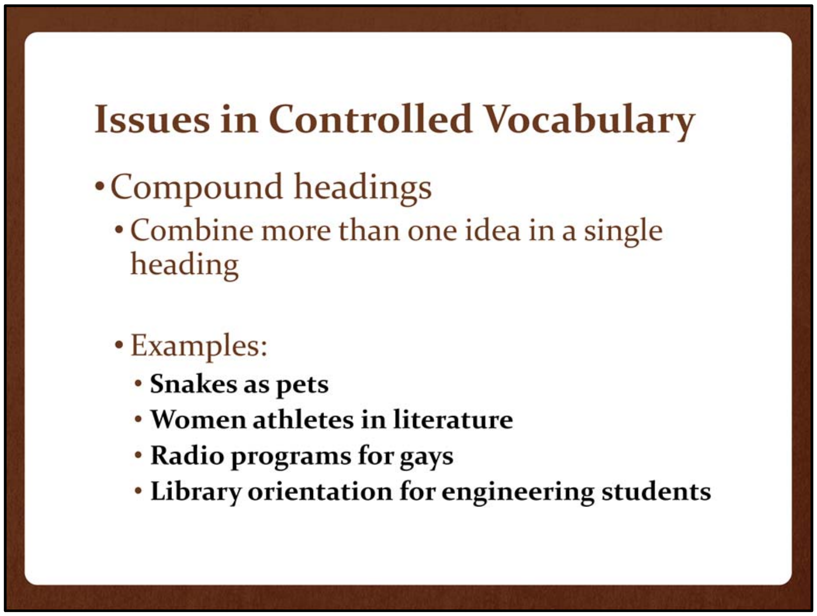
United States—History—Revolution, 1775-1783—Campaigns

UF	United States—History—Revolution, 1775-1783—Campaigns and battles
NT	Bedford (Westchester County, N.Y.)—History—Burning by the British, 1779
	Bennington, Battle of, N.Y., 1777
	...

This is a more complex heading string from LCSH. It contains multiple subdivisions. You can see that the heading is **United States—History—Revolution, 1775-1783—Campaigns**. Every subdivision is set off with a dash, so you can see that there are three of them. The subdivisions bring out topical and chronological aspects of the **History** of the **United States**.

This heading also has a UF for an additional way to phrase the heading. You can see that the UF has *Campaigns and battles* as the final subdivision, instead of **Campaigns**, as in the authorized heading.

It also has a list of narrower terms, of which only two are provided here. These narrower terms are headings for individual campaigns within the revolution. These campaigns can be battles; they can be sieges, and so forth. This brings together all of the military actions that are named, that are established in LCSH. They are all in one place in order to help the user find materials they need.



Issues in Controlled Vocabulary

- Compound headings
  - Combine more than one idea in a single heading
  - Examples:
    - **Snakes as pets**
    - **Women athletes in literature**
    - **Radio programs for gays**
    - **Library orientation for engineering students**

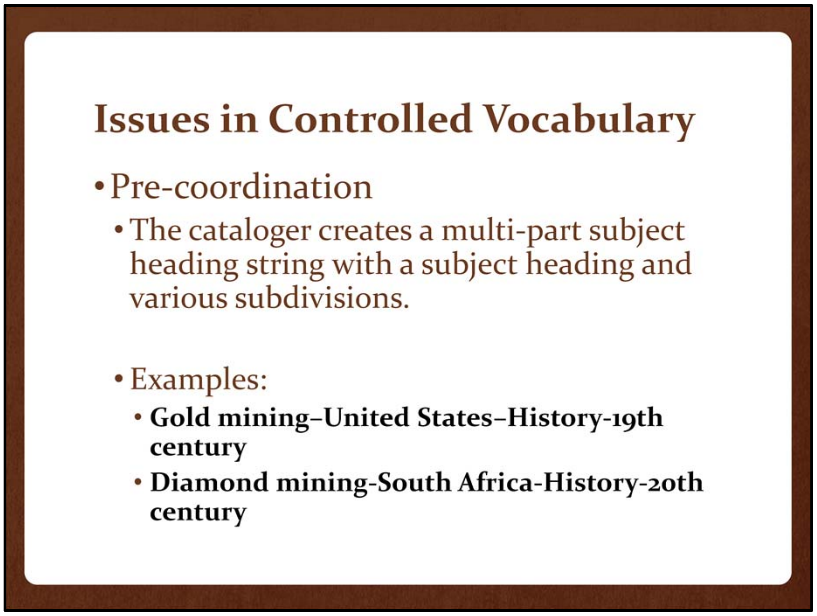
In some controlled vocabularies, such as thesauri, single concepts are the norm; the terms are mostly uncomplicated and very focused.

In other vocabularies, such as subject heading lists, you may find compound headings. This is the idea that two or more concepts may be placed together into a phrase heading – as seen on the slide. In the examples we see **Snakes as pets**. We have a couple of ideas in that single phrase.

Women athletes in literature; **Radio programs for gays**; and **Library orientation for engineering students**.

One has to be careful in a controlled vocabulary, because too much compounding can result in the hierarchies (the broader terms and narrower terms) becoming unclear – a little muddied. Not everything could or should be combined in a phrase heading.

For example, LCSH is probably not going to establish a heading like “Crocheting of novelty potholders” because then they might have to establish headings for other possible products that one could crochet: “Crocheting of baby clothes”, “Crocheting of sweaters”, “Crocheting of doilies”, as well as “Crocheting regular potholders”.



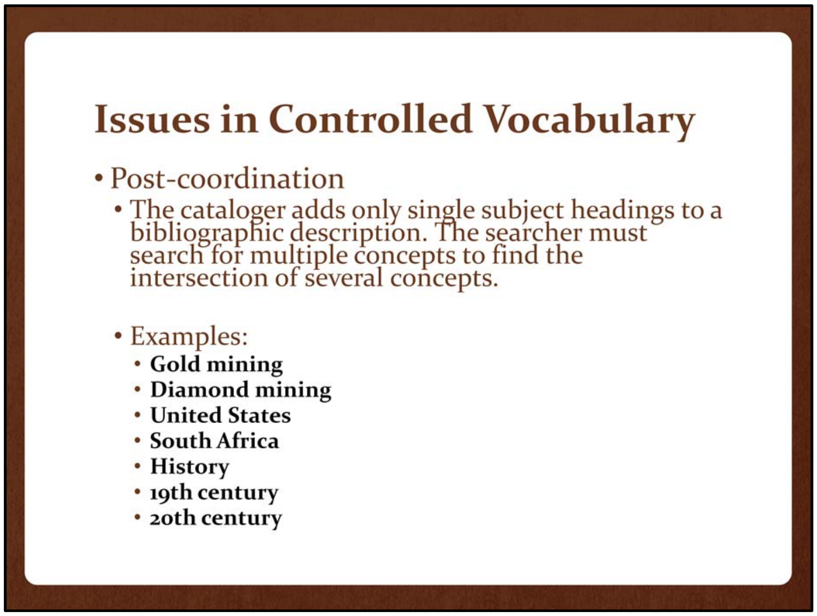
Issues in Controlled Vocabulary

- Pre-coordination
  - The cataloger creates a multi-part subject heading string with a subject heading and various subdivisions.
  - Examples:
    - **Gold mining–United States–History-19th century**
    - **Diamond mining-South Africa-History-20th century**

In addition to compound headings, one might encounter pre-coordinated subject strings in LCSH as well. In pre-coordinated indexing, appropriate terms are chosen and coordinated into subject-subdivision combinations at the time of indexing or cataloging. On the screen you'll see a couple of examples of pre-coordinated strings.

You'll see **Gold mining—United States—History—19th century**. What we have here is a topical subject heading followed by a geographic subdivision, a topical subdivision (**History**), and a chronological period (**19th century**).

In the second, it is the same structure, but with different components in that particular heading string. In each of these cases, what you get from the pre-coordinated string is a certain amount of context.



Issues in Controlled Vocabulary

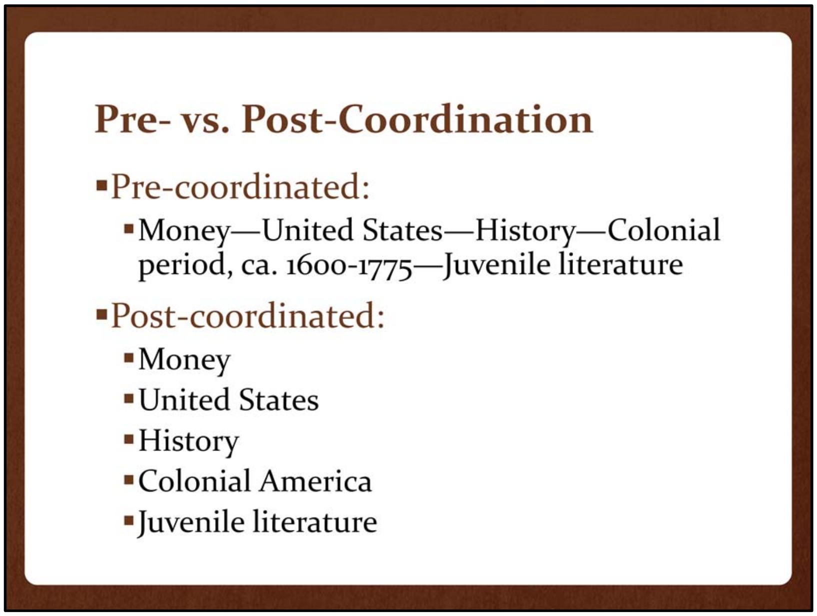
- Post-coordination
  - The cataloger adds only single subject headings to a bibliographic description. The searcher must search for multiple concepts to find the intersection of several concepts.
  - Examples:
    - **Gold mining**
    - **Diamond mining**
    - **United States**
    - **South Africa**
    - **History**
    - **19th century**
    - **20th century**

Post-coordinate indexing, on the other hand, does not provide such context. With post-coordinate indexing, subject concepts are entered as single terms so that users are required to coordinate them. Boolean searching and other advanced techniques are required in order to locate resources on the compound and/or complex subjects in which the searchers are interested.

False associations may easily occur because relationships among terms can be unclear.

If you look at the slide, you can see an example of what happens when terms are post-coordinated instead of being used in pre-coordinated strings. In contrast to the previous slide, in which it was clear that the resource was about gold mining in the United States in the 19th century, in this example it is unclear whether it's about gold mining in the United States in the 19th century, or about diamond mining in South Africa in the 19th century, or diamond mining in the United States. And what century is it?

This is how false drops happen. A post-coordinated system often causes a significant number of false drops, because the context is missing. One might never know exactly what that work is about, without the strings.



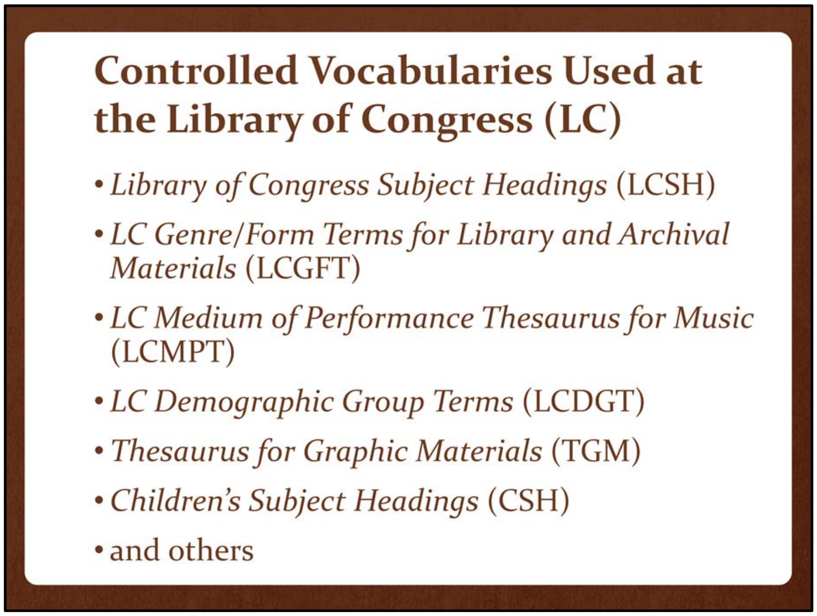
Pre- vs. Post-Coordination

- Pre-coordinated:
  - Money—United States—History—Colonial period, ca. 1600-1775—Juvenile literature
- Post-coordinated:
  - Money
  - United States
  - History
  - Colonial America
  - Juvenile literature

This is another example of pre- and post-coordination. As you can see in the top example which is a pre-coordinated string, it is very clear that the resource is about the history of money during the American colonial period, and it is for children – therefore the subdivision **Juvenile literature**.

On the other hand, with post-coordination, individual terms and phrases (**Money**, **United States**, **History**, **Colonial America**, and **Juvenile literature**) would be assigned to the same resource. It is unclear whether this is a history of juvenile literature, or perhaps a history of the United States during the colonial period.

As you can see, the pre-coordinated strings do provide essential context.



Controlled Vocabularies Used at the Library of Congress (LC)

- *Library of Congress Subject Headings* (LCSH)
- *LC Genre/Form Terms for Library and Archival Materials* (LCGFT)
- *LC Medium of Performance Thesaurus for Music* (LCMPT)
- *LC Demographic Group Terms* (LCDGT)
- *Thesaurus for Graphic Materials* (TGM)
- *Children's Subject Headings* (CSH)
- and others

We've now looked at various types of controlled vocabularies. Multiple controlled vocabularies may be in use within a particular given library. For instance, the Library of Congress uses numerous vocabularies and even different types of vocabularies.

LCSH (or *Library of Congress Subject Headings*) is a subject heading list, as its title implies. *Library of Congress Genre/Form Terms*, and *Medium of Performance Thesaurus* are both true thesauri. So is the *Thesaurus for Graphic Materials*.

All of the vocabularies on your screen are also developed and maintained at the Library of Congress, but the Library of Congress uses vocabularies developed and maintained by other institutions as well. For example, medical subject headings appear in the Library of Congress catalog. Medical subject headings are maintained by the National Library of Medicine.

The Library of Congress also uses the *Art and Architecture Thesaurus* which is maintained by the Getty.