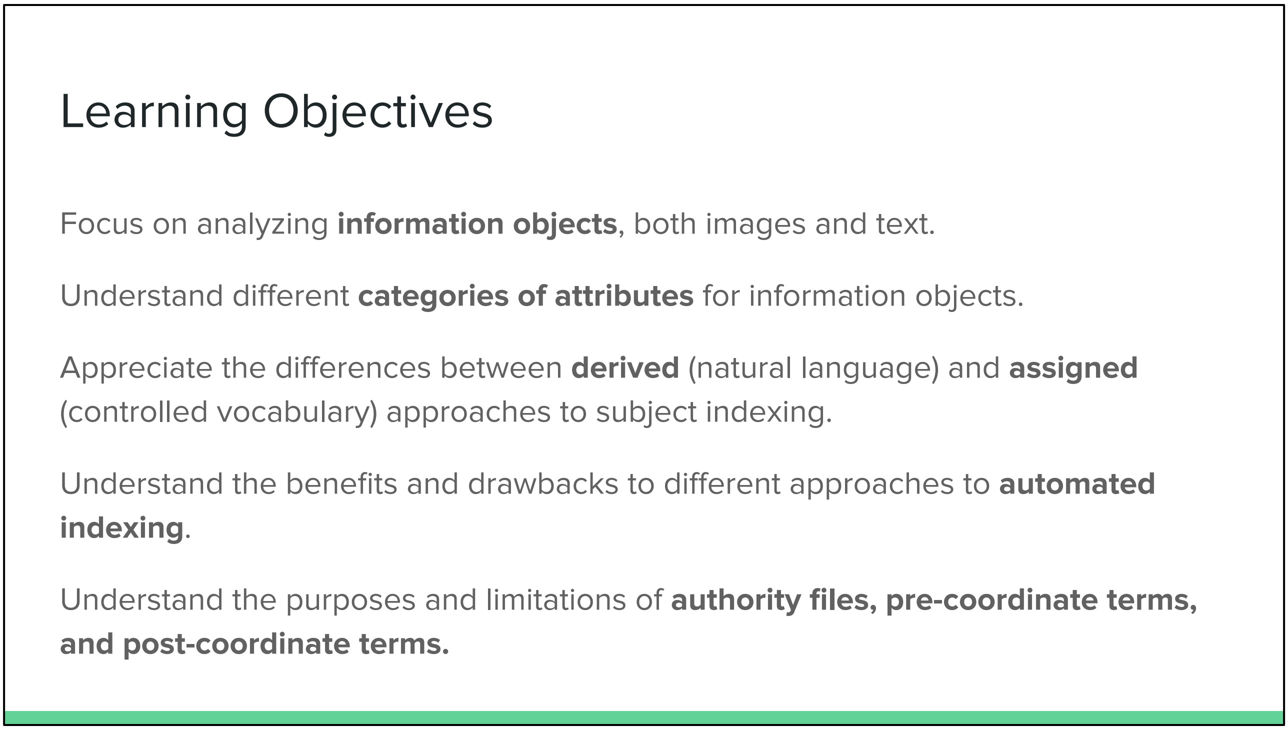
# Subject Indexing: Information Objects and Concepts in Taxonomies and Ontologies

IMT 530
Spring 2019



## Learning Objectives

Focus on analyzing **information objects**, both images and text.

Understand different **categories of attributes** for information objects.

Appreciate the differences between **derived** (natural language) and **assigned** (controlled vocabulary) approaches to subject indexing.

Understand the benefits and drawbacks to different approaches to **automated indexing**.

Understand the purposes and limitations of **authority files, pre-coordinate terms, and post-coordinate terms.**

Note that we've been mostly talking about concepts and their relationships, like beers -> ales -> IPAs. Today we want to think about the other end of the spectrum, individual information objects--this particular bottle of beer. We need to describe the bottle of beer (that's called subject analysis), so we can characterize it and collocate it with similar information objects in a taxonomy or ontology.

So, today we're going to talk about subject analysis and how to use that subject analysis when we build taxonomies and ontologies.

We'll look at Layne's paper on indexing images to understand four different categories of attributes for information objects and explore how that plays out for images. We'll talk about the difference between "aboutness" and "ofness" and the differences and relationships between say a photograph of an object and the object itself. We'll take a brief look at some tools from Google and Microsoft for analyzing images.

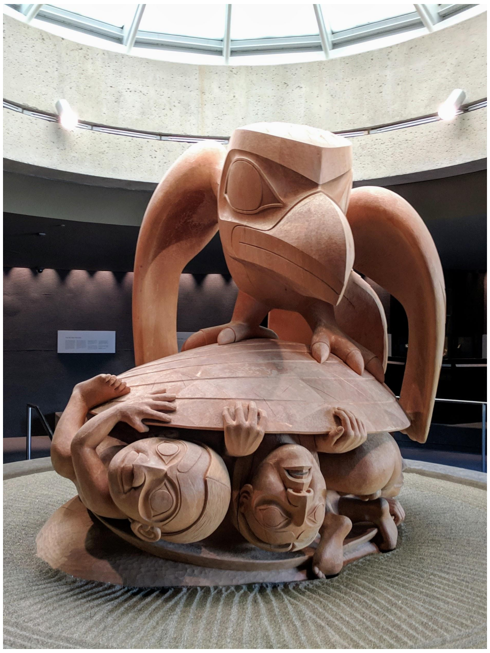
We'll also look at indexing text, both human and machine approaches. When we index text, we can take a derived approach (language that appears naturally in the documents) or an assigned approach (which requires more interpretation and aligns more readily with a taxonomy or ontology with defined terms). We'll look at human approaches to text indexing and machine approaches, including information extraction and auto-categorization.

We'll talk a bit about warrant and concept inclusion. Today we're talking about information objects, how we analyze what they are and what they are about, and how we relate that analysis to taxonomies and ontologies. That relationship part raises a question: Which comes first? The ontology or the information object? Do you base the concepts in your ontology on the objects that you encounter? Or do you fit the objects into your ontology? There's a bit of back-and-forth here and we'll discuss that.

Oh, and we'll talk about how these things affect precision and recall in information retrieval.

Finally, if we have time, I'll share an example from my research.

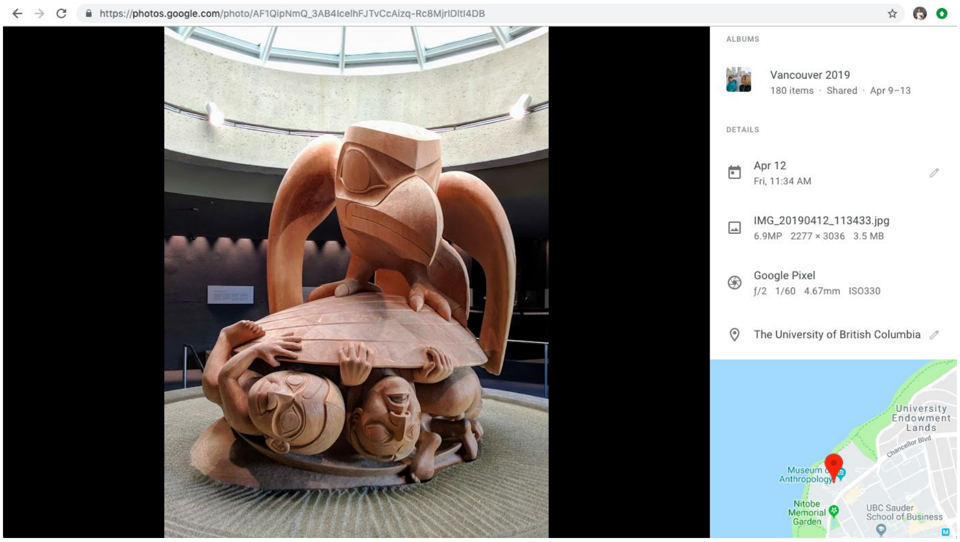
That's a lot to cover. Let's get going!



What can you tell me about this picture?
What is it of?
What is it about?
What else can you tell me about it?

Let’s look at this more carefully with the four categories of attributes from “Some Issues in the Indexing of Images” by Sara Layne, 1994.

# “Biographical” Attributes



https://photos.google.com/photo/AF1QipNmQ_3AB4IcelhFJTvCcAizq-Rc8MjrlDltl4DB

ALBUMS

Vancouver 2019
180 items · Shared · Apr 9–13

DETAILS

Apr 12
Fri, 11:34 AM

IMG_20190412_113433.jpg
6.9MP 2277 × 3036 3.5 MB

Google Pixel
ƒ/2 1/60 4.67mm ISO330

The University of British Columbia

Biographical data is metadata like:

- When and where the picture was taken
- Who took the picture
- Where it has been (Does the digital album count here?) More interesting for say, paintings where we care a lot about provenance, where it came from.
- Title, if it has one
- Maybe some information about how it was edited. I know that this picture was initially landscape and that I adjusted the brightness a bit. Oh, I also created a different version by editing to black and white.

This data tends to be objective, which means that it is administrative.

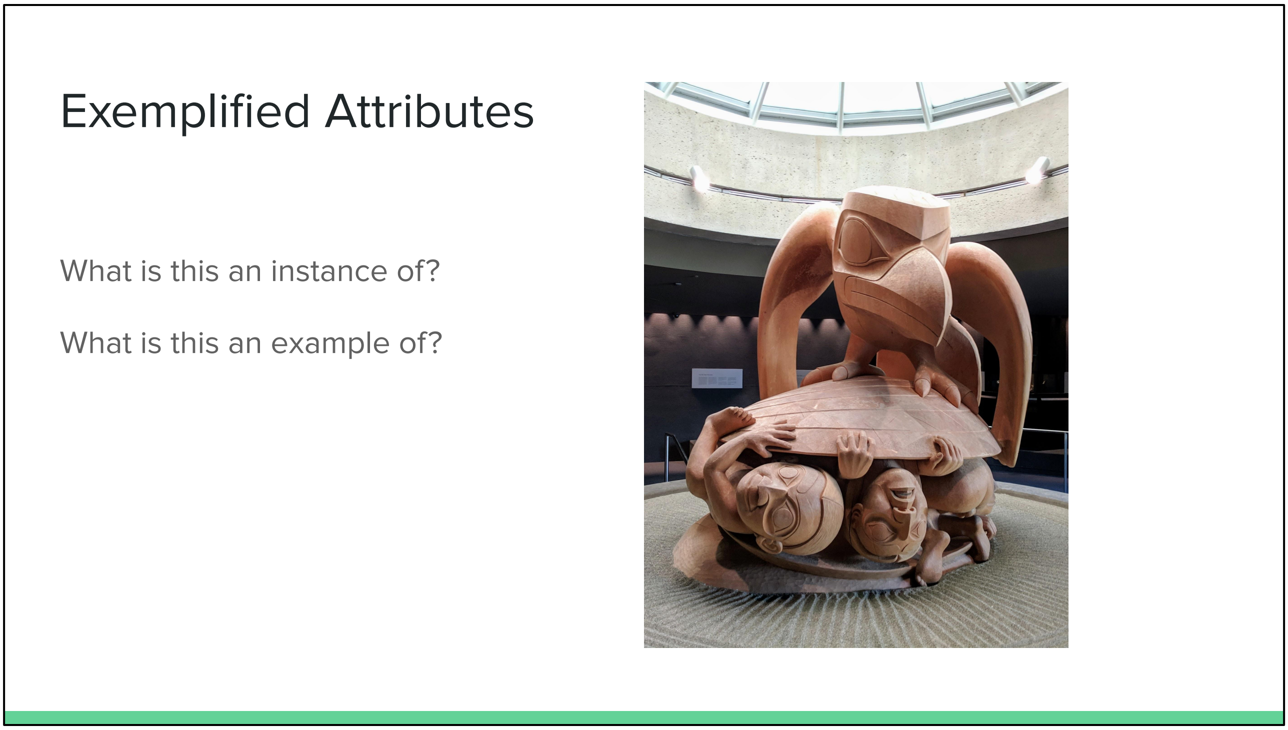
Let’s talk about PoolParty briefly here, specifically the choice between concepts in your taxonomy and attributes in a custom scheme. Would Date WHen Photograph was Taken work better as a concept or an attribute? (Attribute because you have potentially lots and lots of dates and a taxonomy of every data in history isn’t very interesting.) How about location? Depends. How many locations do you have? Maybe it makes sense to have a taxonomy of locations in Vancouver and Museum of Anthropology could be one of those locations.

Note that the metadata from Google Photos is specifically about my photo. We could also pull together biographical data for the actual wood carving, not my photograph. The wood carving is related to my photograph, but, importantly, it is a different information object.

For the actual carving:

- Title: The Raven and the First Men
- Artist: Bill Reid
- Location: Museum of Anthropology at University of British Columbia (It was installed there upon completion.)
- Completed: 1980
- Size: About 8 feet tall

And there’s a whole bunch more biographical data...



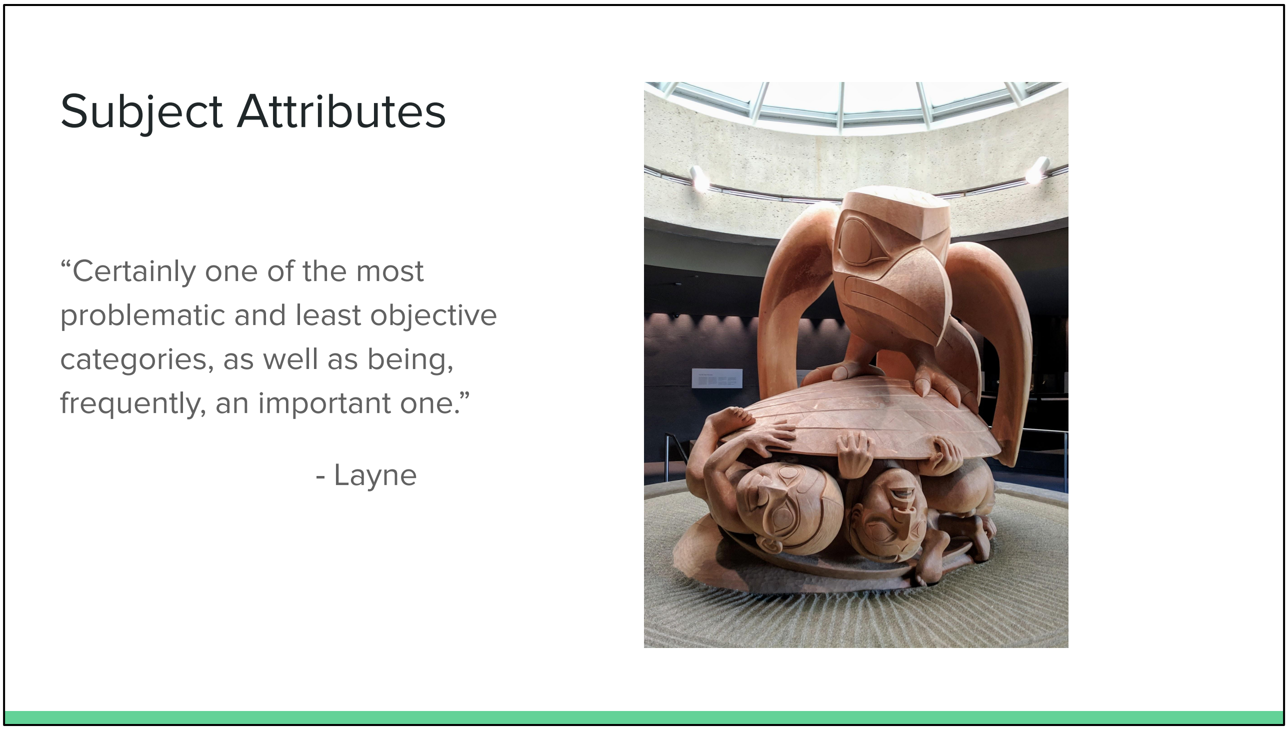
# Exemplified Attributes

What is this an instance of?

What is this an example of?

The photograph is a photograph. We could get more detailed: digital photograph, color photograph.

The artwork is a sculpture. It is made of wood and sand.



## Subject Attributes

> “Certainly one of the most problematic and least objective categories, as well as being, frequently, an important one.”
>
> - Layne

Now we’re firmly into descriptive (subjective) attributes and even more importantly, we’re moving from “ofness” to “aboutness.”

This photograph shows a raven. That’s what it is *of*. It shows a wooden sculpture. That’s what it is *of*.

But it also depicts the origin story of the Haida people. That is what it is *about*. There’s a whole rich set of meaning here that the photograph and the sculpture are about. The sculpture is *about* telling the stories of native peoples. You could argue that the photograph is also *about* my trip to Vancouver, which isn’t readily apparent from looking at the photograph.

Note that subject attributes are much more difficult to derive by just looking at the information object. You likely need to research related information objects to understand aboutness. When you do this, you start to understand richer relationships with other information objects.

Note also that machines are not good at determining subject attributes. Machines have a much easier time identifying what a photograph is *of* than what it is *about*, as we’ll see in a couple of slides.

## Relationship Attributes

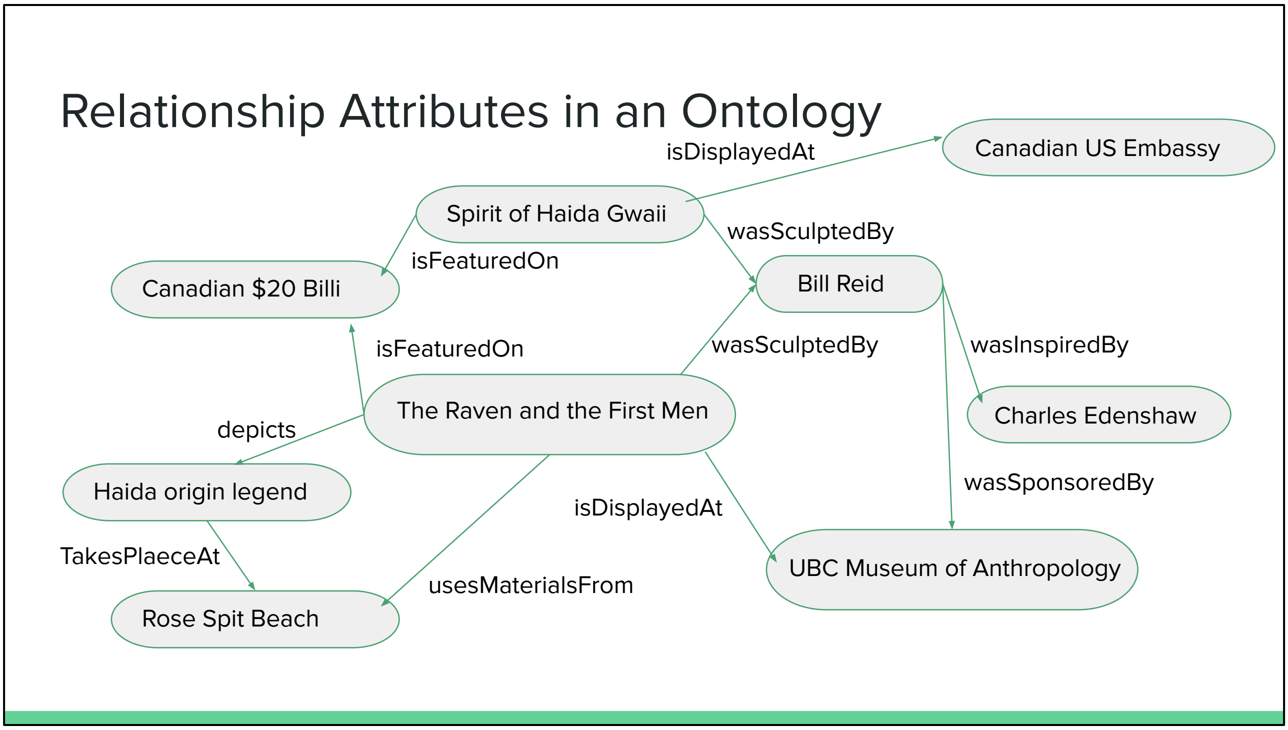
What do we see here?

- Bill Reid, the artist
- A book about the Legend of the Raven
- Rose Pit Beach, the location of the origin story
- A Canadian $20 bill with an image of the sculpture on it
- A small, golden version of the sculpture that was created before the large wooden one. There is also a medium-sized clay version.
- Another famous sculpture by Bill Reid, Spirit of Haida Gwaii

These information objects are all related to the information object that we started with. When we build a taxonomy, we want to reflect these relationships. For example:

The Raven and the First Men [was sculpted by] Bill Reid.
Spirit of Haida Gwaii [was sculpted by] Bill Reid.
The Raven and the First Men [is a] Wooden Sculpture.
The Raven and the First Men [appears on] the Canadian $20 bill.
The Raven and the First Men [is displayed at] the UBC Museum of Anthropology.
etc.

We are using triples here. That’s one way of expressing what we learn about an information object when we perform subject analysis.



# Relationship Attributes in an Ontology

isDisplayedAt

Canadian US Embassy

Spirit of Haida Gwaii

wasSculptedBy

Canadian $20 Billi

isFeaturedOn

Bill Reid

isFeaturedOn

wasSculptedBy

wasInspiredBy

The Raven and the First Men

Charles Edenshaw

depicts

Haida origin legend

wasSponsoredBy

isDisplayedAt

TakesPlaeceAt

UBC Museum of Anthropology

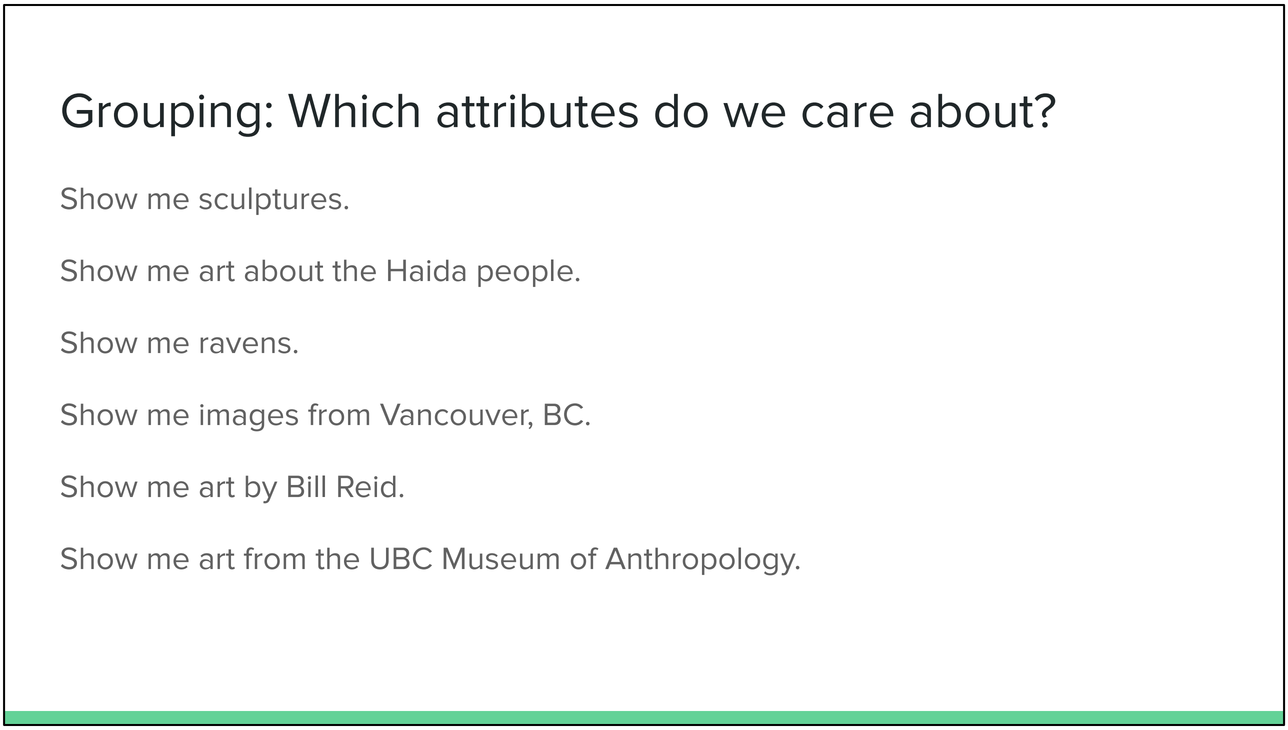
usesMaterialsFrom

Rose Spit Beach

Could also add Bill Reid [wrote book about] haid origin legend.

What else can we add to this?

Quote: “All models are wrong. Some are useful.”



## Grouping: Which attributes do we care about?

Show me sculptures.

Show me art about the Haida people.

Show me ravens.

Show me images from Vancouver, BC.

Show me art by Bill Reid.

Show me art from the UBC Museum of Anthropology.

These are just a few examples of the types of groupings that a user might request. We want to support many different forms of access by supporting many different groupings. We do this by associating metadata with information objects.

Layne: “That is, it is necessary to determine which attributes are needed to provide useful groupings of images; which attributes provide information that is useful once the images are found; and which attributes may, or even should, be left to the searcher or researcher to identify.”

# Concept Inclusion

Is the concept in scope?

Is the concept important?

Is there enough information about the concept?

Do users want and expect the concept to be included?

Note: This doesn’t apply exclusively to text-based information objects. It applies to images, etc. too.

We’ve analyzed our information objects and we know a ton about them. Should everything we know about an information object be a concept? The extreme case of this would be including every single word of a document. We probably don’t want to go that far. We want to be smart about which concepts we include...

From Accidental Taxonomist

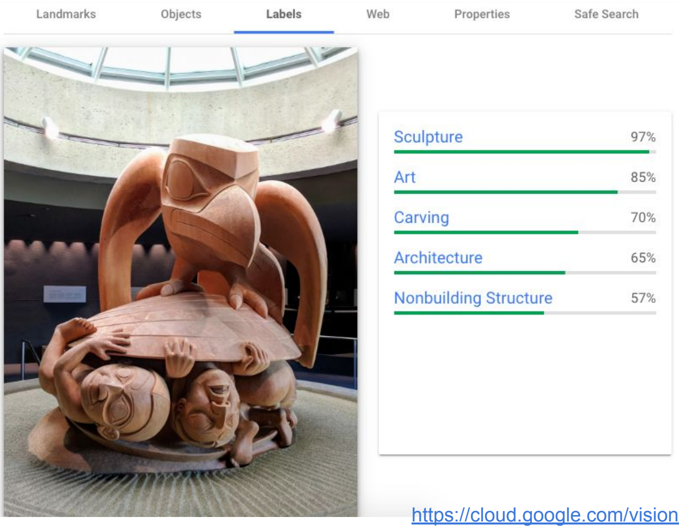
**Scope** is important, you don’t want to get too big. Especially true for a hierarchical taxonomy. Think about the purpose of your taxonomy or ontology.

**Importance** is very subjective, but it again comes back to the purpose of your ontology and what your users care about.

**Enough information:** You might think that a concept is important, but if you don’t have information objects that you can label with that concept, it might not make sense in your taxonomy or ontology.

**User warrant:** This is particularly important for cultural groups like native peoples, who have historically had other groups decide what should be included for them. Think about their search use cases. What groups of information objects would they care about. Think back to “Show me art by Bill Reid” or “Show me  ravens.”

# Machine Analysis



| | |
|---|---|
| Sculpture | 97% |
| Art | 85% |
| Carving | 70% |
| Architecture | 65% |
| Nonbuilding Structure | 57% |

https://cloud.google.com/vision/docs/drag-and-drop

**Go to the webpage itself to show the different types of data produced. Map to Laynes attributes.**

Note that the machine is better at “ofness” data than “aboutness” data.

It does recognize that this is a famous landmark, though, and can locate it on a map. That’s pretty cool.

Important: This process can be done programmatically in bulk with lots and lots of pictures analyzed very quickly. This type of analysis is used for Google Image search, Google Photos, and similar technologies by other companies.

Question: Do you think that these labels come from a controlled vocabulary?

# Indexing Texts

**Similar to images:**

- “Biographical” attributes
- Type of work (photograph, sculpture, book, etc.)
- Subject attributes: “Ofness” and “aboutness”
- Relationships between information objects
- Grouping like items

**Different:**

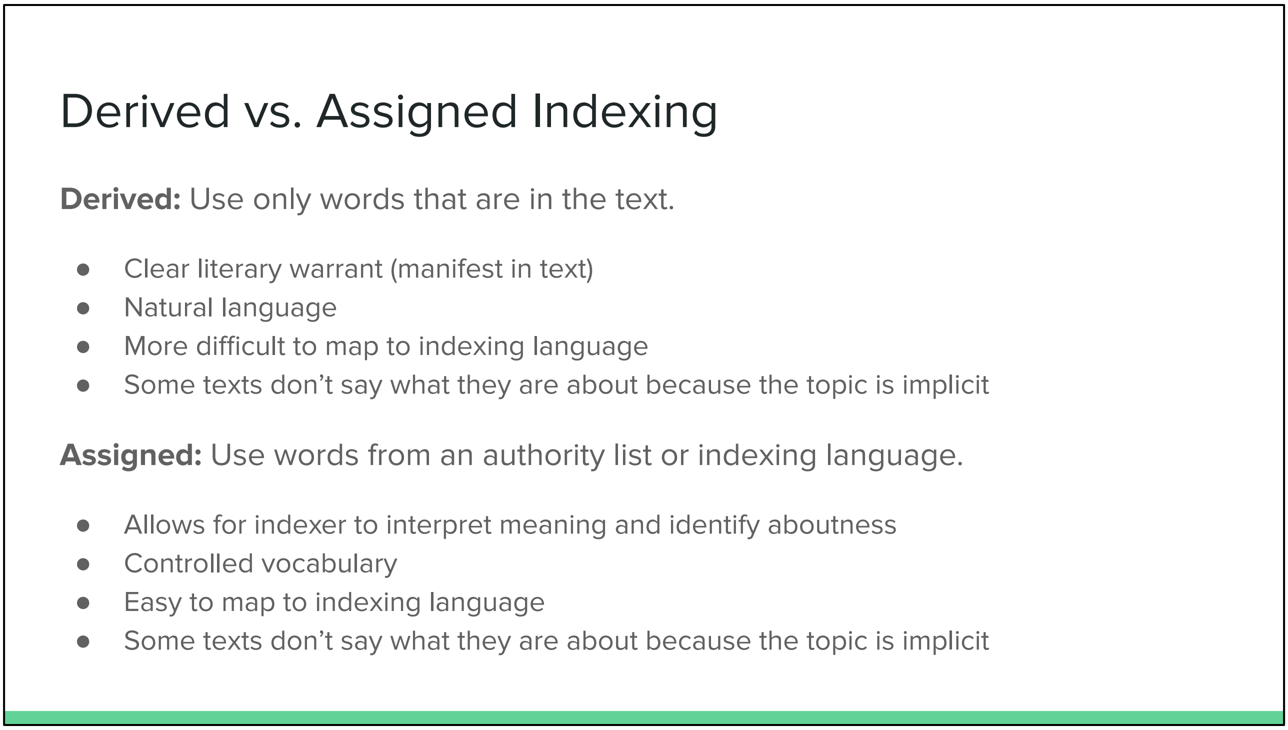
We have words! (Sometimes lots of words.)

Okay, let’s shift gears from images to text. We have some significant similarities:

- We still identify “biographical”attributes: publication date, title, author, word count, edition number, etc.
- We still identify the type of work. Is it a photograph? Sculpture? Book? Journal article? Web page? Etc.
- We still care about “ofness” and “aboutness.” We might be able to get a sense of “ofness” from word frequency, but “aboutness” is still a challenge that requires more in-depth analysis.
- We still care about relationships. We want to know how texts are related, just like images.
- We still care about grouping. We want to collocate similar texts.

But the big difference is that we now have words! (Maybe lots and lots of words.) And words are what we use as labels or values for attributes. We don’t need to look at a picture and interpret a bird to be a raven. We (or a machine) can just read the word raven.

Also different: We have a longer and more advanced tradition of indexing texts -- think libraries. And we have conventions like frontmatter, “about this book,” standard metadata, chapters, etc.



# Derived vs. Assigned Indexing

**Derived:** Use only words that are in the text.

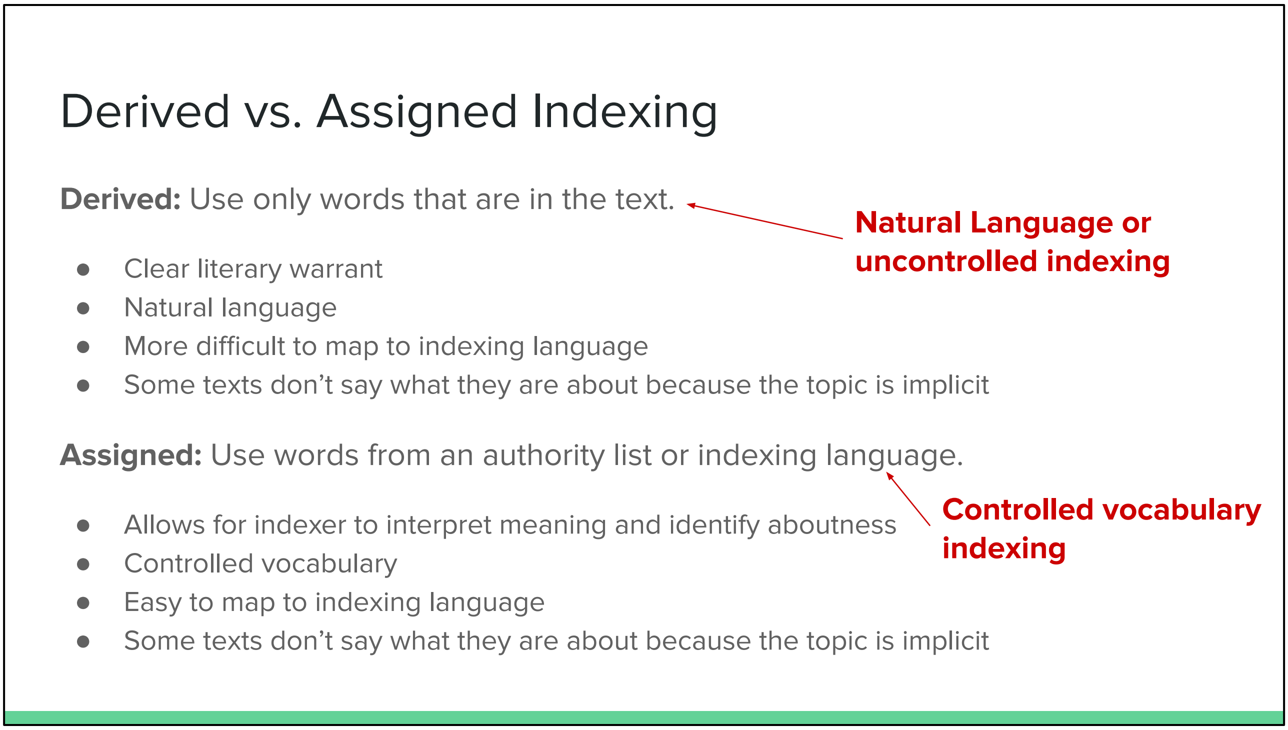
- Clear literary warrant (manifest in text)
- Natural language
- More difficult to map to indexing language
- Some texts don’t say what they are about because the topic is implicit

**Assigned:** Use words from an authority list or indexing language.

- Allows for indexer to interpret meaning and identify aboutness
- Controlled vocabulary
- Easy to map to indexing language
- Some texts don’t say what they are about because the topic is implicit

Because we have words, we can choose to use directly them or to interpret them.

Derived: Use only the words that are explicitly in the text. The indexer is deriving the terms directly from the text.
Assigned: The indexer assigns terms based on their interpretation of the text.



# Derived vs. Assigned Indexing

**Derived:** Use only words that are in the text.

Natural Language or uncontrolled indexing

- Clear literary warrant
- Natural language
- More difficult to map to indexing language
- Some texts don't say what they are about because the topic is implicit

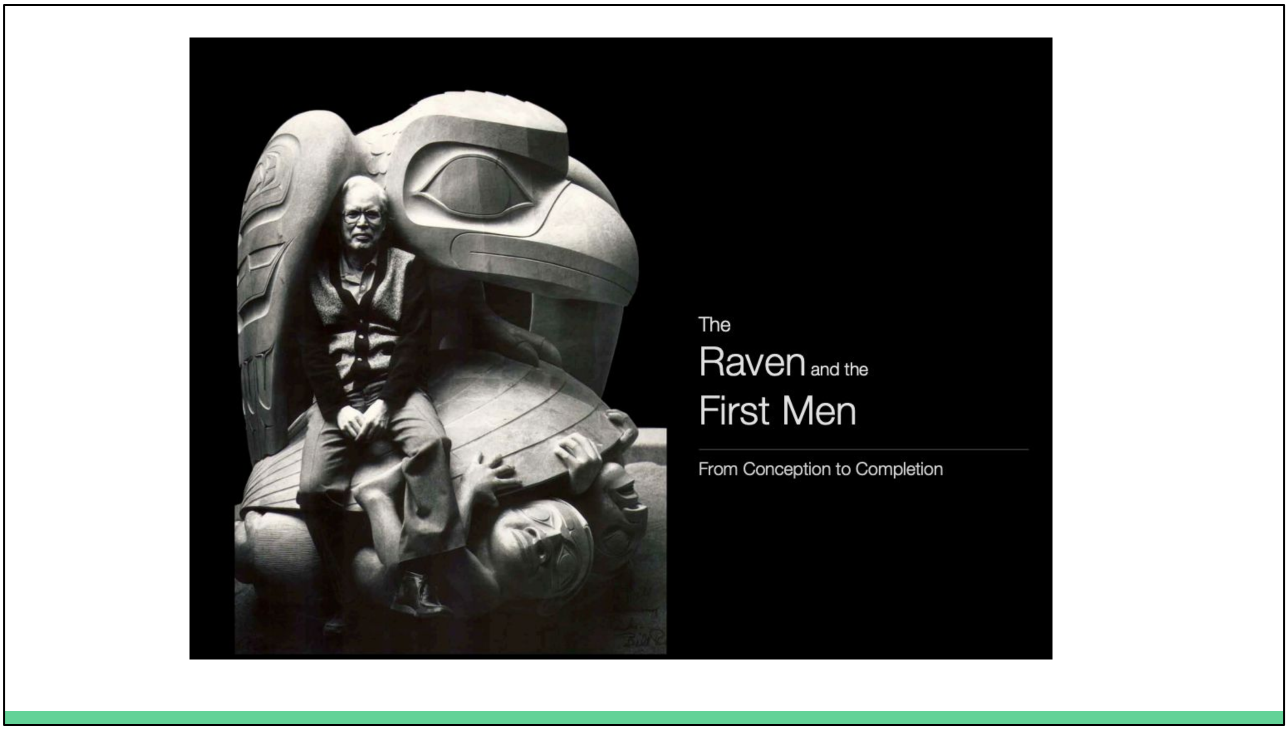
**Assigned:** Use words from an authority list or indexing language.

Controlled vocabulary indexing

- Allows for indexer to interpret meaning and identify aboutness
- Controlled vocabulary
- Easy to map to indexing language
- Some texts don't say what they are about because the topic is implicit

Oh, by the way, sometimes we call this difference uncontrolled vs. controlled indexing.

Note that this distinction doesn't really make sense for images because for the most part we cannot derive words that are directly manifest in a photograph or work of art.



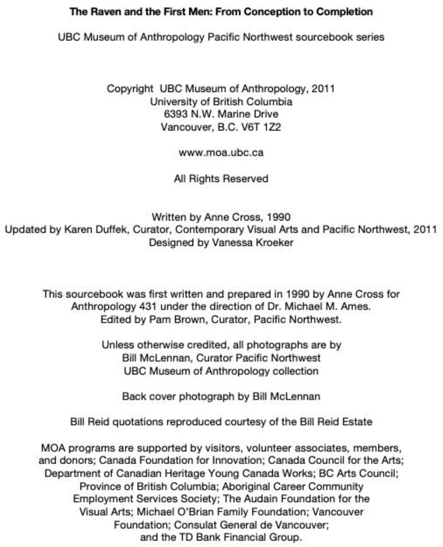
The Raven and the First Men

From Conception to Completion

Let’s do some subject analysis for a text about the Raven and the First Men.

# Biographical Attributes

Frontmatter is a key place to find this information in texts, especially books.



**The Raven and the First Men: From Conception to Completion**

UBC Museum of Anthropology Pacific Northwest sourcebook series

Copyright  UBC Museum of Anthropology, 2011
University of British Columbia
6393 N.W. Marine Drive
Vancouver, B.C. V6T 1Z2

www.moa.ubc.ca

All Rights Reserved

Written by Anne Cross, 1990
Updated by Karen Duffek, Curator, Contemporary Visual Arts and Pacific Northwest, 2011
Designed by Vanessa Kroeker

This sourcebook was first written and prepared in 1990 by Anne Cross for
Anthropology 431 under the direction of Dr. Michael M. Ames.
Edited by Pam Brown, Curator, Pacific Northwest.

Unless otherwise credited, all photographs are by
Bill McLennan, Curator Pacific Northwest
UBC Museum of Anthropology collection

Back cover photograph by Bill McLennan

Bill Reid quotations reproduced courtesy of the Bill Reid Estate

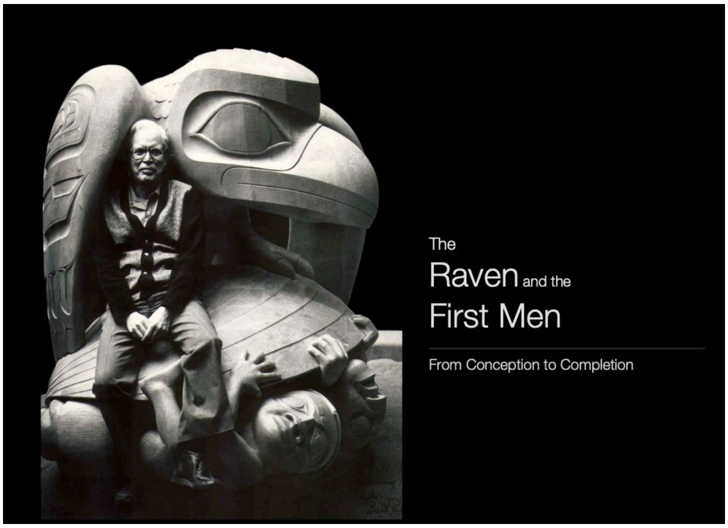
MOA programs are supported by visitors, volunteer associates, members, and donors; Canada Foundation for Innovation; Canada Council for the Arts; Department of Canadian Heritage Young Canada Works; BC Arts Council; Province of British Columbia; Aboriginal Career Community Employment Services Society; The Audain Foundation for the Visual Arts; Michael O'Brian Family Foundation; Vancouver Foundation; Consulat General de Vancouver; and the TD Bank Financial Group.

“Biographical” attributes: let’s look in particular at the front matter.

## Exemplified Attributes

What is this an instance of?

What is this an example of?

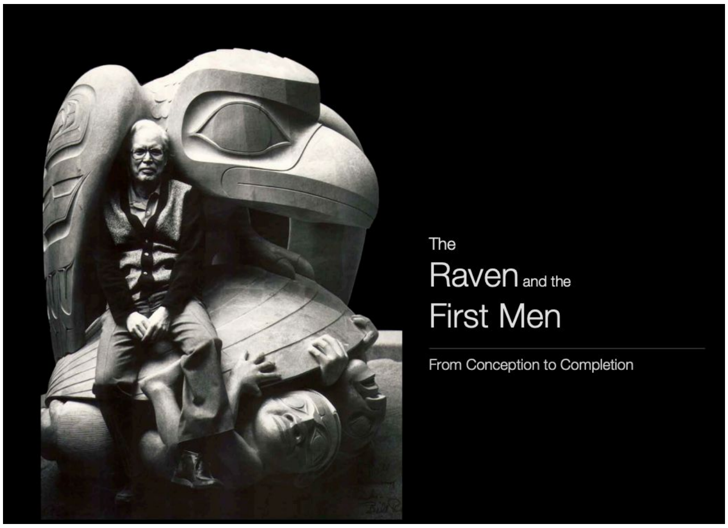


The publishers call it a “source book.”

# Subject Attributes

Let’s consider:

- Derived vs. Assigned
- Scope
- Importance
- Available information
- User expectations



Derived: Lots of terms are there:

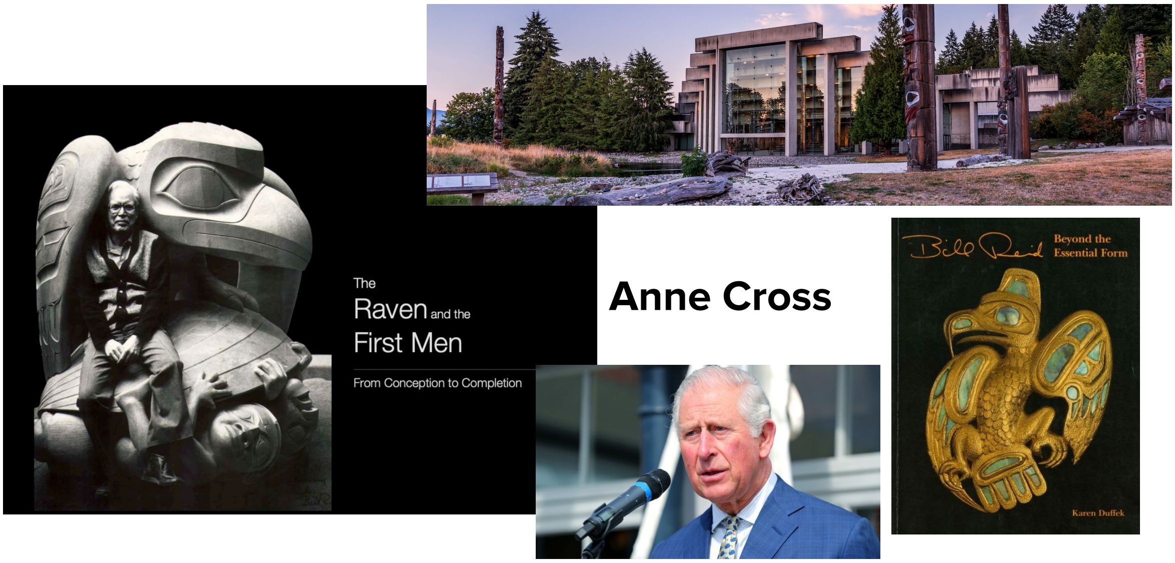
- Bill Reid
- Sculpture
- Art
- Raven
- Haida
- Legend

Some aren’t:

- Native
- Sand

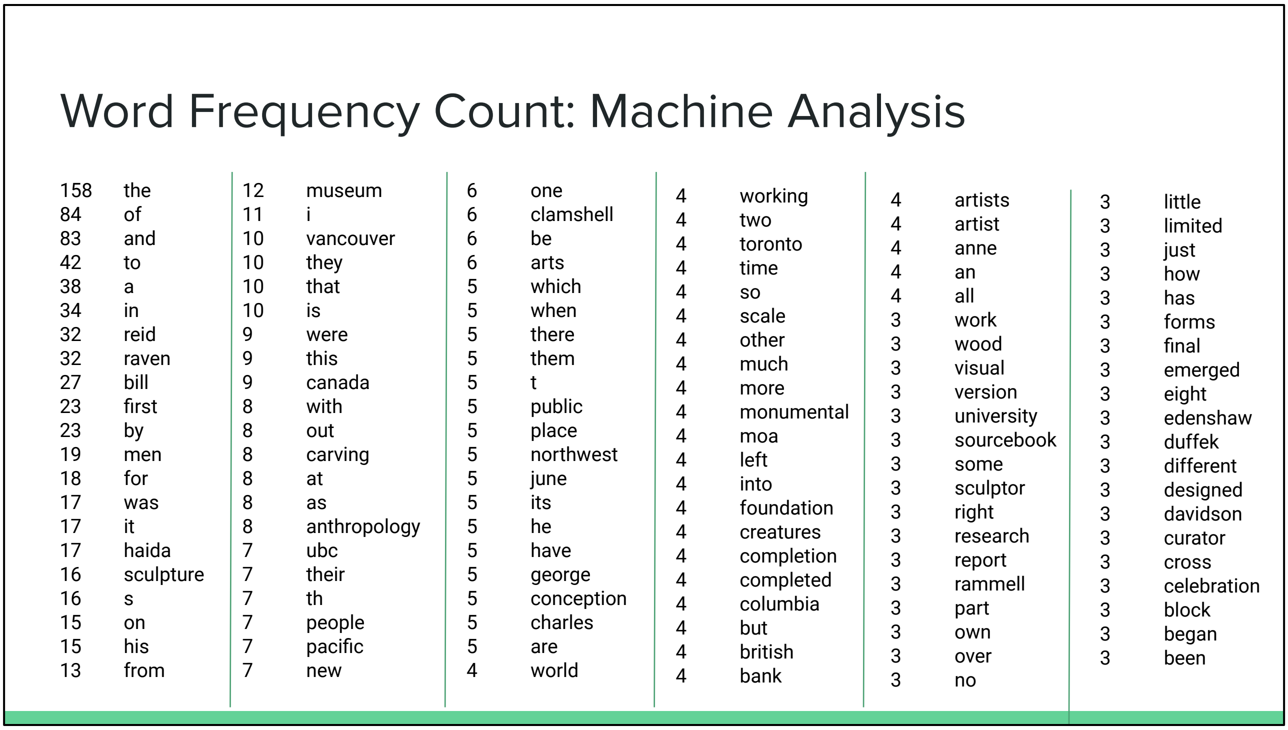
Do we use these terms? Depends….

# Relationship Attributes



**Anne Cross**

Produced by UBC MOA
Written by Anne Cross
Cites “Bill Reid: Beyond the Essential Form” by Karen Duffek
The text is about the sculpture itself, so that’s another relationship...



## Word Frequency Count: Machine Analysis

| | |
|---|---|
| 158 | the |
| 84 | of |
| 83 | and |
| 42 | to |
| 38 | a |
| 34 | in |
| 32 | reid |
| 32 | raven |
| 27 | bill |
| 23 | first |
| 23 | by |
| 19 | men |
| 18 | for |
| 17 | was |
| 17 | it |
| 17 | haida |
| 16 | sculpture |
| 16 | s |
| 15 | on |
| 15 | his |
| 13 | from |
| 12 | museum |
| 11 | i |
| 10 | vancouver |
| 10 | they |
| 10 | that |
| 10 | is |
| 9 | were |
| 9 | this |
| 9 | canada |
| 8 | with |
| 8 | out |
| 8 | carving |
| 8 | at |
| 8 | as |
| 8 | anthropology |
| 7 | ubc |
| 7 | their |
| 7 | th |
| 7 | people |
| 7 | pacific |
| 7 | new |
| 6 | one |
| 6 | clamshell |
| 6 | be |
| 6 | arts |
| 5 | which |
| 5 | when |
| 5 | there |
| 5 | them |
| 5 | t |
| 5 | public |
| 5 | place |
| 5 | northwest |
| 5 | june |
| 5 | its |
| 5 | he |
| 5 | have |
| 5 | george |
| 5 | conception |
| 5 | charles |
| 5 | are |
| 4 | world |
| 4 | working |
| 4 | two |
| 4 | toronto |
| 4 | time |
| 4 | so |
| 4 | scale |
| 4 | other |
| 4 | much |
| 4 | more |
| 4 | monumental |
| 4 | moa |
| 4 | left |
| 4 | into |
| 4 | foundation |
| 4 | creatures |
| 4 | completion |
| 4 | completed |
| 4 | columbia |
| 4 | but |
| 4 | british |
| 4 | bank |
| 4 | artists |
| 4 | artist |
| 4 | anne |
| 4 | an |
| 4 | all |
| 3 | work |
| 3 | wood |
| 3 | visual |
| 3 | version |
| 3 | university |
| 3 | sourcebook |
| 3 | some |
| 3 | sculptor |
| 3 | right |
| 3 | research |
| 3 | report |
| 3 | rammell |
| 3 | part |
| 3 | own |
| 3 | over |
| 3 | no |
| 3 | little |
| 3 | limited |
| 3 | just |
| 3 | how |
| 3 | has |
| 3 | forms |
| 3 | final |
| 3 | emerged |
| 3 | eight |
| 3 | edenshaw |
| 3 | duffek |
| 3 | different |
| 3 | designed |
| 3 | davidson |
| 3 | curator |
| 3 | cross |
| 3 | celebration |
| 3 | block |
| 3 | began |
| 3 | been |

First, there's a ton of words below that I couldn't fit on the slide.

There are a bunch of stop words that we'd typically filter out: the, of, an, to, a, in, by for, was, it, on, I, etc.

How well do these words convey "aboutness"?

We can use this type of analysis on an individual document to get some sense of what the document is about.

But we can also use it on a full corpus (lots of documents) to understand the important concepts in that corpus.

Typically this type of machine analysis is used as a starting point, not a complete analysis.

I used a "dumb" web tool (http://www.writewords.org.uk/word_count.asp) to grab these words. Smarter tools might differentiate between different parts of the text. The title and chapter and section headings might be weighted as particularly important. The frontmatter might be treated differently to make it easier to mine biographical data.

Question: This analysis might give us some hints about term inclusion, especially if we're taking a derived indexing approach. Can this type of analysis help us establish relationship attributes?

Words that didn’t fit on the slide:

| | |
|---|---|
| 3 | beams |
| 3 | art |
| 3 | about |
| 2 | young |
| 2 | you |
| 2 | york |
| 2 | years |
| 2 | written |
| 2 | worlds |
| 2 | works |
| 2 | while |
| 2 | we |
| 2 | way |
| 2 | visitors |
| 2 | visit |
| 2 | very |
| 2 | us |
| 2 | until |
| 2 | ulli |
| 2 | these |
| 2 | then |
| 2 | teaching |
| 2 | strong |
| 2 | steltzer |
| 2 | sort |
| 2 | something |
| 2 | society |
| 2 | small |
| 2 | size |
| 2 | shell |
| 2 | shadbolt |
| 2 | series |
| 2 | see |
| 2 | saw |
| 2 | royal |
| 2 | rotunda |
| 2 | reg |
| 2 | programs |
| 2 | produced |
| 2 | prince |

| | |
|---|---|
| 2 | press |
| 2 | prepared |
| 2 | photo |
| 2 | past |
| 2 | original |
| 2 | or |
| 2 | number |
| 2 | not |
| 2 | norris |
| 2 | my |
| 2 | months |
| 2 | model |
| 2 | miniature |
| 2 | min |
| 2 | michael |
| 2 | mclennan |
| 2 | mcintyre |
| 2 | long |
| 2 | like |
| 2 | legend |
| 2 | leave |
| 2 | karen |
| 2 | jim |
| 2 | interpretation |
| 2 | inspired |
| 2 | humans |
| 2 | highness |
| 2 | having |
| 2 | hart |
| 2 | had |
| 2 | gwaii |
| 2 | guujaaw |
| 2 | guests |
| 2 | great |
| 2 | get |
| 2 | found |
| 2 | feet |
| 2 | favourite |
| 2 | far |
| 2 | extraordinary |
| 2 | experience |
| 2 | eventually |
| 2 | european |
| 2 | dried |
| 2 | douglas |

| | |
|---|---|
| 2 | doris |
| 2 | director |
| 2 | dedication |
| 2 | de |
| 2 | day |
| 2 | curiosity |
| 2 | cultures |
| 2 | creating |
| 2 | created |
| 2 | create |
| 2 | council |
| 2 | correct |
| 2 | contemporary |
| 2 | community |
| 2 | century |
| 2 | celebrated |
| 2 | cbc |
| 2 | carver |
| 2 | carved |
| 2 | career |
| 2 | canadian |
| 2 | boxwood |
| 2 | born |
| 2 | before |
| 2 | become |
| 2 | bc |
| 2 | back |
| 2 | assistants |
| 2 | assistance |
| 2 | ancestors |
| 2 | american |
| 2 | achievements |
| 2 | above |
| 1 | z |
| 1 | yet |
| 1 | yellow |
| 1 | www |
| 1 | writes |
| 1 | worldview |
| 1 | workable |
| 1 | wonderful |
| 1 | within |
| 1 | wings |
| 1 | wider |
| 1 | who |

| | |
|---|---|
| 1 | where |
| 1 | west |
| 1 | went |
| 1 | well |
| 1 | waved |
| 1 | wasn |
| 1 | washington |
| 1 | walter |
| 1 | waking |
| 1 | wade |
| 1 | w |
| 1 | volunteer |
| 1 | visions |
| 1 | view |
| 1 | victoria |
| 1 | vibrant |
| 1 | variety |
| 1 | vanessa |
| 1 | v |
| 1 | using |
| 1 | use |
| 1 | updated |
| 1 | up |
| 1 | unveiling |
| 1 | unveiled |
| 1 | unscramble |
| 1 | unless |
| 1 | unique |
| 1 | uninvolved |
| 1 | unformed |
| 1 | understand |
| 1 | under |
| 1 | twists |
| 1 | twenties |
| 1 | twelve |
| 1 | tv |
| 1 | turned |
| 1 | traditions |
| 1 | toward |
| 1 | touch |
| 1 | tool |
| 1 | too |
| 1 | tony |
| 1 | ton |
| 1 | today |

| | |
|---|---|
| 1 | timidly |
| 1 | thrusting |
| 1 | through |
| 1 | three |
| 1 | though |
| 1 | thompson |
| 1 | third |
| 1 | think |
| 1 | thing |
| 1 | than |
| 1 | terry |
| 1 | tells |
| 1 | technologists |
| 1 | techniques |
| 1 | td |
| 1 | take |
| 1 | suppose |
| 1 | supporter |
| 1 | supported |
| 1 | supernatural |
| 1 | suits |
| 1 | suddenly |
| 1 | studio |
| 1 | studied |
| 1 | strange |
| 1 | story |
| 1 | stick |
| 1 | stepped |
| 1 | steals |
| 1 | state |
| 1 | st |
| 1 | spit |
| 1 | spent |
| 1 | spectacular |
| 1 | special |
| 1 | speaking |
| 1 | soon |
| 1 | skylight |
| 1 | sky |
| 1 | skin |
| 1 | skidegate |
| 1 | singers |
| 1 | since |
| 1 | simulate |
| 1 | silversmith |

| | |
|---|---|
| 1 | silence |
| 1 | shed |
| 1 | shape |
| 1 | seven |
| 1 | setting |
| 1 | services |
| 1 | september |
| 1 | self |
| 1 | section |
| 1 | seattle |
| 1 | sea |
| 1 | scurried |
| 1 | sculptures |
| 1 | sculptors |
| 1 | screened |
| 1 | scrambled |
| 1 | scottish |
| 1 | sciences |
| 1 | says |
| 1 | row |
| 1 | round |
| 1 | roughed |
| 1 | rot |
| 1 | rose |
| 1 | role |
| 1 | robert |
| 1 | rights |
| 1 | returning |
| 1 | responses |
| 1 | reserved |
| 1 | resemblance |
| 1 | reproduced |
| 1 | representational |
| 1 | reported |
| 1 | renowned |
| 1 | renewal |
| 1 | remove |
| 1 | remarked |
| 1 | released |
| 1 | referring |
| 1 | redcedar |
| 1 | reasons |
| 1 | really |
| 1 | real |
| 1 | reached |

| | |
|---|---|
| 1 | rayonier |
| 1 | ravenand |
| 1 | radio |
| 1 | quotations |
| 1 | quite |
| 1 | quiet |
| 1 | pushed |
| 1 | provocative |
| 1 | province |
| 1 | proved |
| 1 | protruding |
| 1 | project |
| 1 | progress |
| 1 | programming |
| 1 | profess |
| 1 | process |
| 1 | presentation |
| 1 | preparation |
| 1 | power |
| 1 | potential |
| 1 | position |
| 1 | point |
| 1 | plans |
| 1 | pivotal |
| 1 | piece |
| 1 | pictures |
| 1 | picture |
| 1 | photographs |
| 1 | photograph |
| 1 | personality |
| 1 | personalities |
| 1 | patrons |
| 1 | partly |
| 1 | parkinson |
| 1 | paradoxes |
| 1 | pam |
| 1 | pale |
| 1 | page |
| 1 | overwhelming |
| 1 | overcome |
| 1 | overcame |
| 1 | overall |
| 1 | our |
| 1 | otherwise |
| 1 | others |

| | |
|---|---|
| 1 | origin |
| 1 | order |
| 1 | opposite |
| 1 | opportunities |
| 1 | opening |
| 1 | open |
| 1 | ontario |
| 1 | once |
| 1 | old |
| 1 | often |
| 1 | off |
| 1 | o |
| 1 | ny |
| 1 | note |
| 1 | nightmare |
| 1 | nephew |
| 1 | needed |
| 1 | necessary |
| 1 | nearly |
| 1 | narrated |
| 1 | naked |
| 1 | n |
| 1 | mythical |
| 1 | myth |
| 1 | moving |
| 1 | moves |
| 1 | movement |
| 1 | moved |
| 1 | move |
| 1 | motion |
| 1 | mother |
| 1 | most |
| 1 | mortifee |
| 1 | moment |
| 1 | modern |
| 1 | mists |
| 1 | mins |
| 1 | mine |
| 1 | meters |
| 1 | menil |
| 1 | members |
| 1 | measuring |
| 1 | measure |
| 1 | maternal |
| 1 | masterpiece |

| | |
|---|---|
| 1 | marine |
| 1 | many |
| 1 | mankind |
| 1 | major |
| 1 | made |
| 1 | m |
| 1 | lower |
| 1 | lots |
| 1 | lot |
| 1 | logistics |
| 1 | log |
| 1 | ll |
| 1 | lineage |
| 1 | light |
| 1 | life |
| 1 | legged |
| 1 | legacy |
| 1 | least |
| 1 | later |
| 1 | last |
| 1 | larger |
| 1 | large |
| 1 | laminated |
| 1 | l |
| 1 | kuramoto |
| 1 | kroeker |
| 1 | koppers |
| 1 | koerner |
| 1 | kiln |
| 1 | ken |
| 1 | journey |
| 1 | join |
| 1 | jewellery |
| 1 | issued |
| 1 | introduced |
| 1 | intertwines |
| 1 | international |
| 1 | interaction |
| 1 | institution |
| 1 | instead |
| 1 | installed |
| 1 | installation |
| 1 | insist |
| 1 | innovation |
| 1 | initial |

| | |
|---|---|
| 1 | industrialist |
| 1 | individuality |
| 1 | individualistic |
| 1 | indians |
| 1 | inadvertence |
| 1 | impractical |
| 1 | immensity |
| 1 | immediately |
| 1 | image |
| 1 | illustrious |
| 1 | icons |
| 1 | ibm |
| 1 | human |
| 1 | hovering |
| 1 | holding |
| 1 | history |
| 1 | himself |
| 1 | him |
| 1 | hesitant |
| 1 | heritage |
| 1 | her |
| 1 | heir |
| 1 | heads |
| 1 | harper |
| 1 | hard |
| 1 | half |
| 1 | hair |
| 1 | haidas |
| 1 | guess |
| 1 | grows |
| 1 | growing |
| 1 | group |
| 1 | grey |
| 1 | greatest |
| 1 | grandfather |
| 1 | grained |
| 1 | good |
| 1 | goldsmithing |
| 1 | glossy |
| 1 | glecoff |
| 1 | gladstone |
| 1 | given |
| 1 | gesture |
| 1 | german |
| 1 | general |

| | |
|---|---|
| 1 | gary |
| 1 | gallery |
| 1 | further |
| 1 | free |
| 1 | frame |
| 1 | four |
| 1 | form |
| 1 | forintek |
| 1 | force |
| 1 | fluttered |
| 1 | flat |
| 1 | fit |
| 1 | finishing |
| 1 | fine |
| 1 | finding |
| 1 | financial |
| 1 | featured |
| 1 | feathers |
| 1 | father |
| 1 | family |
| 1 | fall |
| 1 | extent |
| 1 | extend |
| 1 | exploration |
| 1 | expansion |
| 1 | exhibitions |
| 1 | except |
| 1 | events |
| 1 | estate |
| 1 | essential |
| 1 | erickson |
| 1 | enlarging |
| 1 | enjoyment |
| 1 | ended |
| 1 | employment |
| 1 | emotions |
| 1 | edited |
| 1 | early |
| 1 | earlier |
| 1 | dynamic |
| 1 | drive |
| 1 | dramatic |
| 1 | dr |
| 1 | down |
| 1 | donors |

| | |
|---|---|
| 1 | donated |
| 1 | doing |
| 1 | doesn |
| 1 | distinguished |
| 1 | disease |
| 1 | discovering |
| 1 | discover |
| 1 | directions |
| 1 | direction |
| 1 | dimensional |
| 1 | difficult |
| 1 | died |
| 1 | didn |
| 1 | did |
| 1 | diameter |
| 1 | design |
| 1 | descendent |
| 1 | depicting |
| 1 | depicted |
| 1 | department |
| 1 | delight |
| 1 | deliberately |
| 1 | defects |
| 1 | dedicated |
| 1 | deal |
| 1 | dan |
| 1 | culture |
| 1 | crowd |
| 1 | crept |
| 1 | credited |
| 1 | creative |
| 1 | crane |
| 1 | cover |
| 1 | courtesy |
| 1 | course |
| 1 | could |
| 1 | copyright |
| 1 | contributed |
| 1 | contemplation |
| 1 | contained |
| 1 | consulat |
| 1 | constantly |
| 1 | connections |
| 1 | concerned |
| 1 | concept |

| | |
|---|---|
| 1 | composition |
| 1 | complicated |
| 1 | completely |
| 1 | commissioned |
| 1 | collections |
| 1 | collection |
| 1 | coaxed |
| 1 | coast |
| 1 | closely |
| 1 | cliffs |
| 1 | clay |
| 1 | claude |
| 1 | characteristic |
| 1 | change |
| 1 | certainly |
| 1 | centred |
| 1 | celebrate |
| 1 | cedar |
| 1 | caution |
| 1 | carvings |
| 1 | carefully |
| 1 | capture |
| 1 | canadians |
| 1 | campus |
| 1 | came |
| 1 | callipers |
| 1 | ca |
| 1 | c |
| 1 | build |
| 1 | brown |
| 1 | brothers |
| 1 | bronze |
| 1 | bringing |
| 1 | bringhurst |
| 1 | bringer |
| 1 | brian |
| 1 | blackness |
| 1 | black |
| 1 | biography |
| 1 | bibliography |
| 1 | beyond |
| 1 | beloved |
| 1 | believe |
| 1 | being |
| 1 | beforehand |

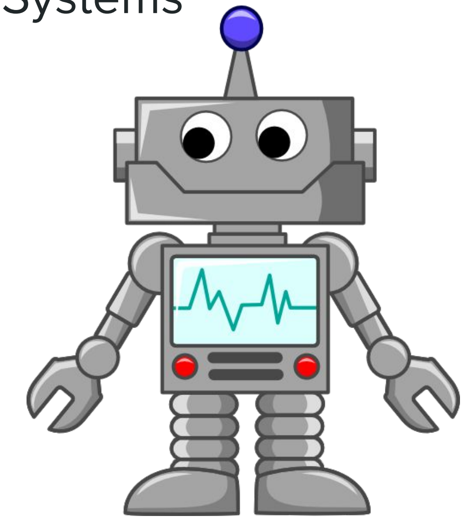
| | |
|---|---|
| 1 | because |
| 1 | became |
| 1 | beauty |
| 1 | beautiful |
| 1 | beak |
| 1 | beach |
| 1 | base |
| 1 | b |
| 1 | awful |
| 1 | awed |
| 1 | audain |
| 1 | attraction |
| 1 | attitude |
| 1 | attempt |
| 1 | associates |
| 1 | arthur |
| 1 | arrive |
| 1 | arrested |
| 1 | around |
| 1 | area |
| 1 | architecture |
| 1 | architectural |
| 1 | architect |
| 1 | april |
| 1 | approximately |
| 1 | applying |
| 1 | appendages |
| 1 | any |
| 1 | anuu |
| 1 | anticipation |
| 1 | another |
| 1 | announcer |
| 1 | anatomy |
| 1 | ames |
| 1 | amazing |
| 1 | am |
| 1 | always |
| 1 | also |
| 1 | already |
| 1 | alone |
| 1 | air |
| 1 | ahead |
| 1 | ages |
| 1 | afflicted |
| 1 | affirm |

| | |
|---|---|
| 1 | adjacent |
| 1 | adelaide |
| 1 | addressing |
| 1 | actually |
| 1 | active |
| 1 | action |
| 1 | achieving |
| 1 | according |
| 1 | acclaimed |
| 1 | accident |
| 1 | aboriginal |

## When to Use Automated Indexing Systems

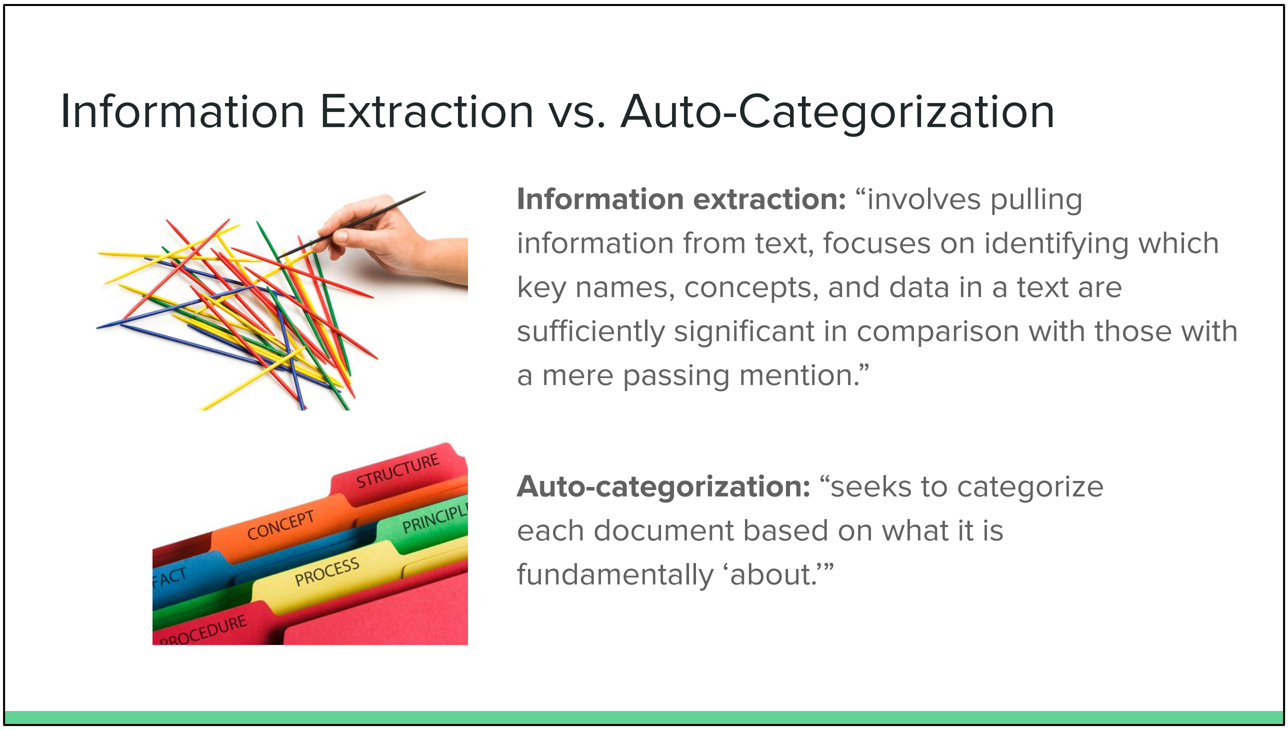
- Large corpus of documents
- Quickly changing content
- Time-critical information, like current events
- Consistent document types of well-structured content
- Content within one subject area
- Text-only content

Note: Even if you use automated indexing, you probably want humans involved in the process at some point.



This is from the Accidental Taxonomist

A very large number of documents, which would require multiple human indexers and would be costly to index.
Content that changes quickly and perhaps unpredictably
A need for speed in indexing, such as for time-critical information, current awareness, or news
Relatively common document types or formats or pretagged (structured) content types
Content related to a relatively uniform subject area or a single industry (so there is less ambiguity of terms)
Text content only (although a few technologies can identify digital video and audio data)
A corporate culture that is more comfortable with investing in externally purchased technology than in hiring, training, and managing human indexers



## Information Extraction vs. Auto-Categorization

**Information extraction:** “involves pulling information from text, focuses on identifying which key names, concepts, and data in a text are sufficiently significant in comparison with those with a mere passing mention.”

**Auto-categorization:** “seeks to categorize each document based on what it is fundamentally ‘about.’”

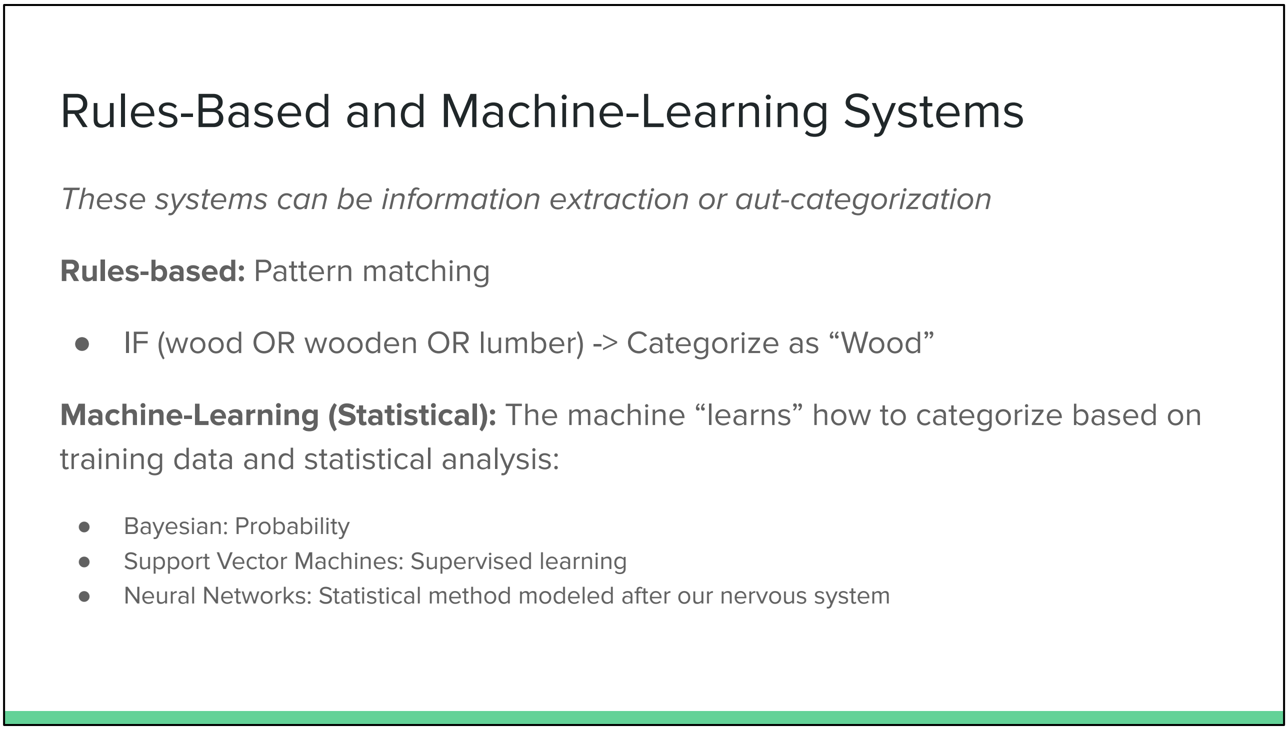
Information extraction is basically a smarter version of the “dumb” word counter that we saw two slides ago.

- About identifying key terms (names, concepts, data) -- Accidental Taxonomist compares it to book indexing.
- Not about organizing or structuring information. Not about relationships.
- Smarter than “dumb”: Eliminating stop words, stemming and synonyms, emphasis on titles and headings, summary statements or “about this work”
- But still very much “ofness” and no “aboutness.” and still more about DERIVED indexing.
- Also called entity extraction. Example: I worked on a natural language processing project last summer. We aimed to extract dates from huge piles of government records. This proved difficult because dates were formatted inconsistently and other numbers, like House of Representatives bill numbers -- looked a bit like dates. We needed to narrow our definition of what we were looking for to succeed. A very narrow way to do entity extraction is to provide a list of entities that you want to match, although there’s usually some effort around synonyms and stemming. Note that we didn’t even get to more interesting questions in our project, like relationships between dates in different documents.

Auto-categorization is more difficult and more sophisticated.

- Does better with structured content than unstructured content because it can make some more assumptions about meaning.
- Usually tries to fit with the categories in an existing taxonomy. Might takes

- steps toward building.
- Not like a book index with lots and lots of key terms. More likely a few key summary terms, like when you see keywords listed for an academic paper. Think ASSIGNED INDEXING.
- Accidental taxonomist is somewhat down on these techniques, suggesting them only for small corpora and with human intervention. I think that they have developed as machine learning has progressed significantly in recent years.



## Rules-Based and Machine-Learning Systems

*These systems can be information extraction or aut-categorization*

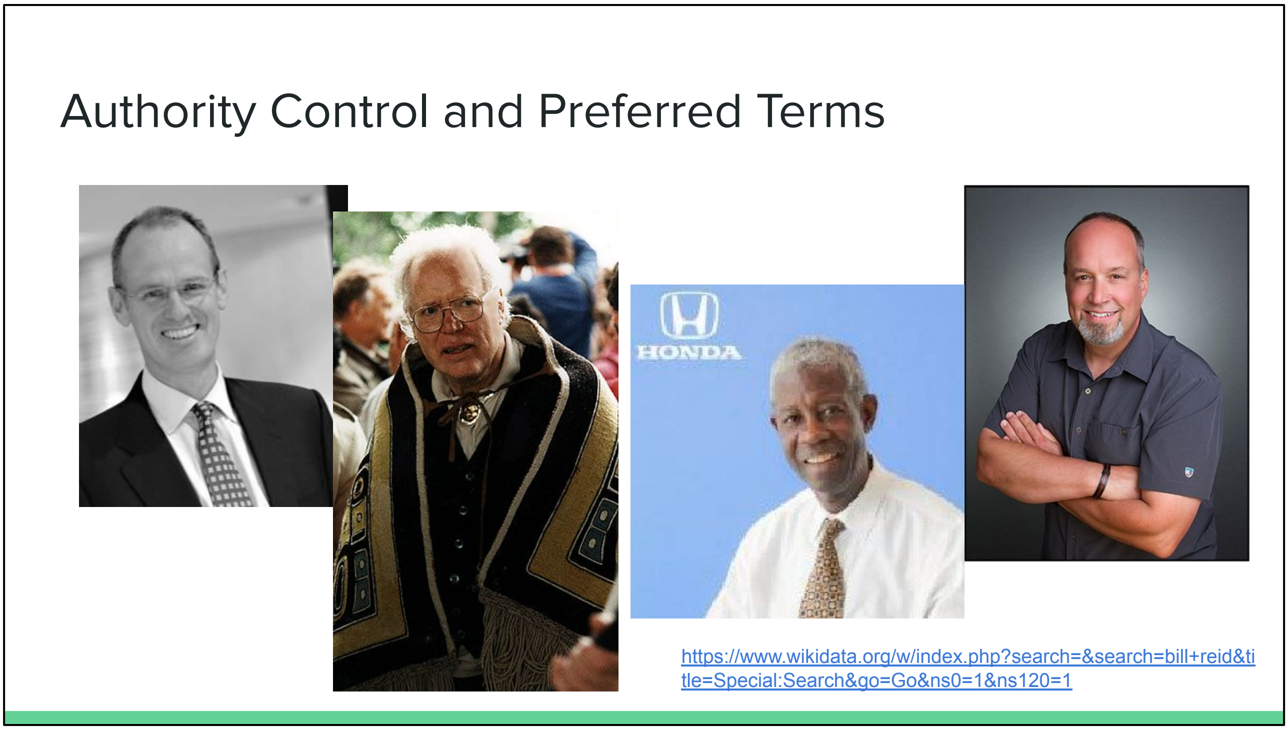
**Rules-based:** Pattern matching

- IF (wood OR wooden OR lumber) -> Categorize as “Wood”

**Machine-Learning (Statistical):** The machine “learns” how to categorize based on training data and statistical analysis:

- Bayesian: Probability
- Support Vector Machines: Supervised learning
- Neural Networks: Statistical method modeled after our nervous system

Key concept for statistical and machine learning



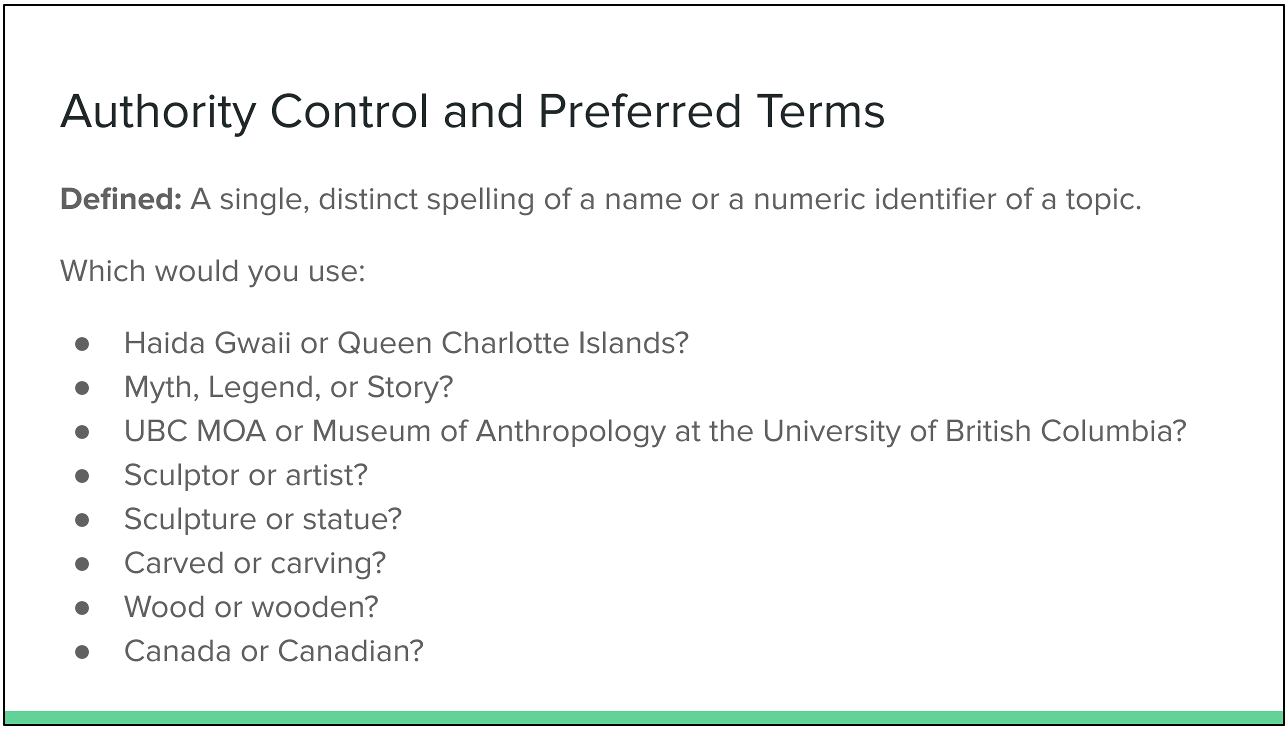
# Authority Control and Preferred Terms

https://www.wikidata.org/w/index.php?search=&search=bill+reid&title=Special:Search&go=Go&ns0=1&ns120=1

Who is Bill Reid?

As far as I know, all of these people are named Bill Reid. We need a distinct (or unique) way to identify a person. We can do this a few ways:

Bill Reid (artist)
Bill Reid (1920-1998)
**Bill Reid** (Q615962) - Follow wikidata link



## Authority Control and Preferred Terms

**Defined:** A single, distinct spelling of a name or a numeric identifier of a topic.

Which would you use:

- Haida Gwaii or Queen Charlotte Islands?
- Myth, Legend, or Story?
- UBC MOA or Museum of Anthropology at the University of British Columbia?
- Sculptor or artist?
- Sculpture or statue?
- Carved or carving?
- Wood or wooden?
- Canada or Canadian?

Haida Gwaii or QCI are near synonyms with significant cultural meaning and history behind the different names. Haida Gwaii is almost certainly a more appropriate choice for a taxonomy or ontology about First Nations art of the Haida people.

Myth, Legend, or Story are also near synonyms with some baggage associated with the words. We might want to go to literary warrant here. Legend is used in a heading in our document, so that might be the right choice. Myth and Story are also used.

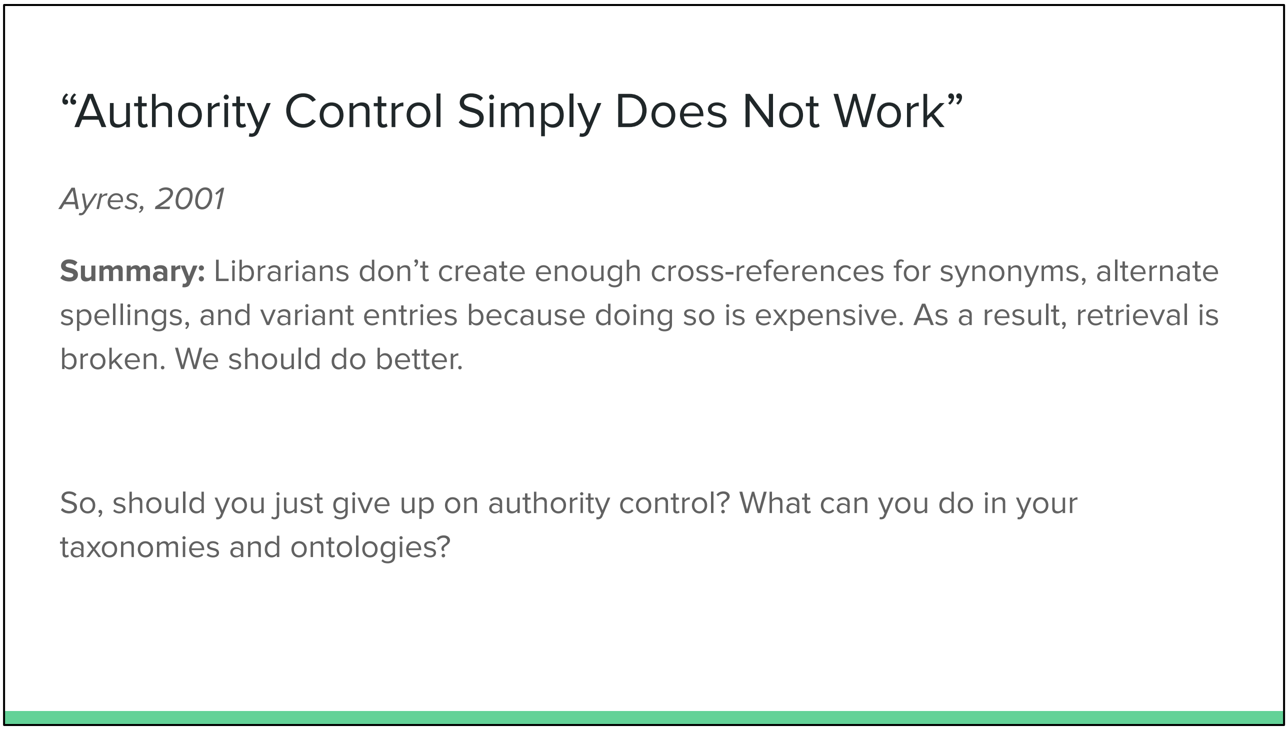
UBC MOA vs. spelling it out might be a choice of different displays for different audiences, but the fully spelled out name is the least ambiguous. There are not well-known initialisms.

Sculptor or artist fill Bill Reid? This is a broader term vs. narrower term distinction. Reid created art other than sculpture, but he specifically sculpted The Artist and the Raven. There's an argument for including both of these terms (artists could have other narrower terms like painter, etc.) and applying them appropriately on a case-base-case basis.

Sculpture or statue are near synonyms. Sculpture is much more heavily used in our text and makes it more clear how the art was created, so I'd go with sculpture.

Carved vs carving is a stemming choice. Important to be consistent with other similar terms. So, make the same choice for painting vs. painted.

Also, How do differentiate between the small golden version of *The Raven and the First Men* and the large wooden one?



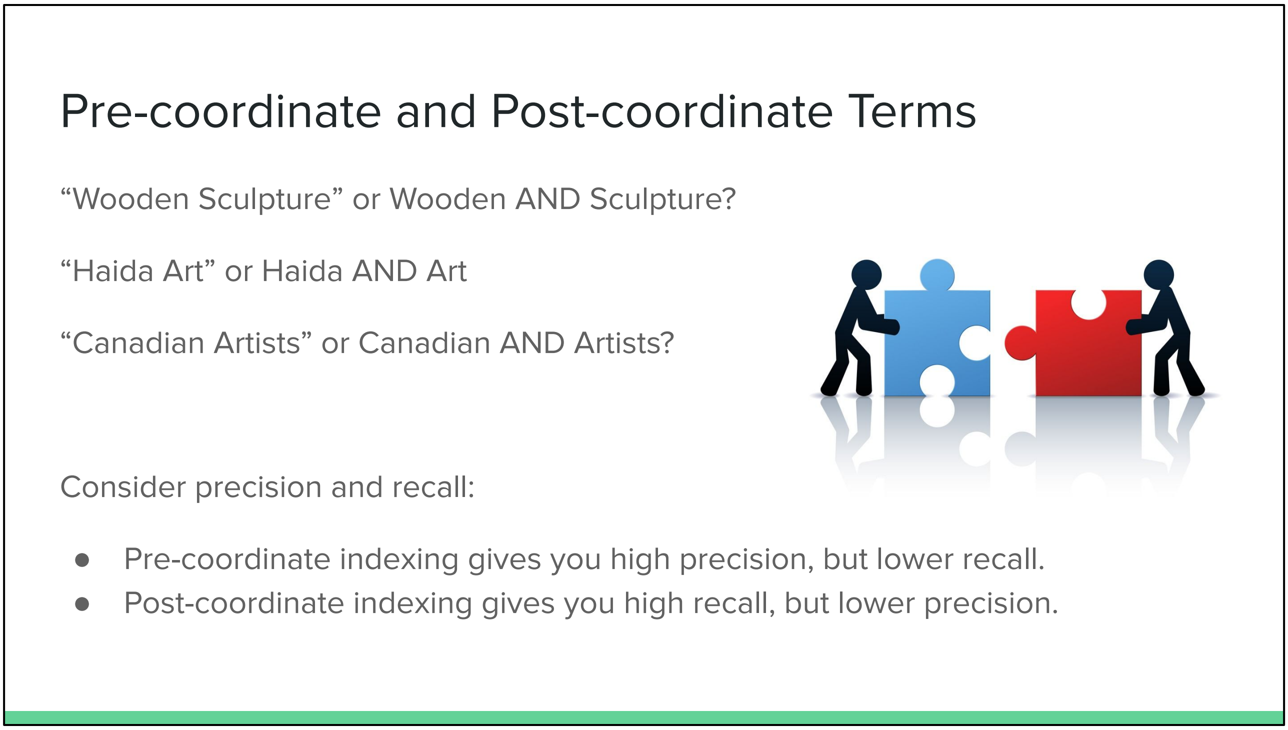
# “Authority Control Simply Does Not Work”

*Ayres, 2001*

**Summary:** Librarians don’t create enough cross-references for synonyms, alternate spellings, and variant entries because doing so is expensive. As a result, retrieval is broken. We should do better.

So, should you just give up on authority control? What can you do in your taxonomies and ontologies?

Make good use of preferred and alternate terms!



## Pre-coordinate and Post-coordinate Terms

“Wooden Sculpture” or Wooden AND Sculpture?

“Haida Art” or Haida AND Art

“Canadian Artists” or Canadian AND Artists?

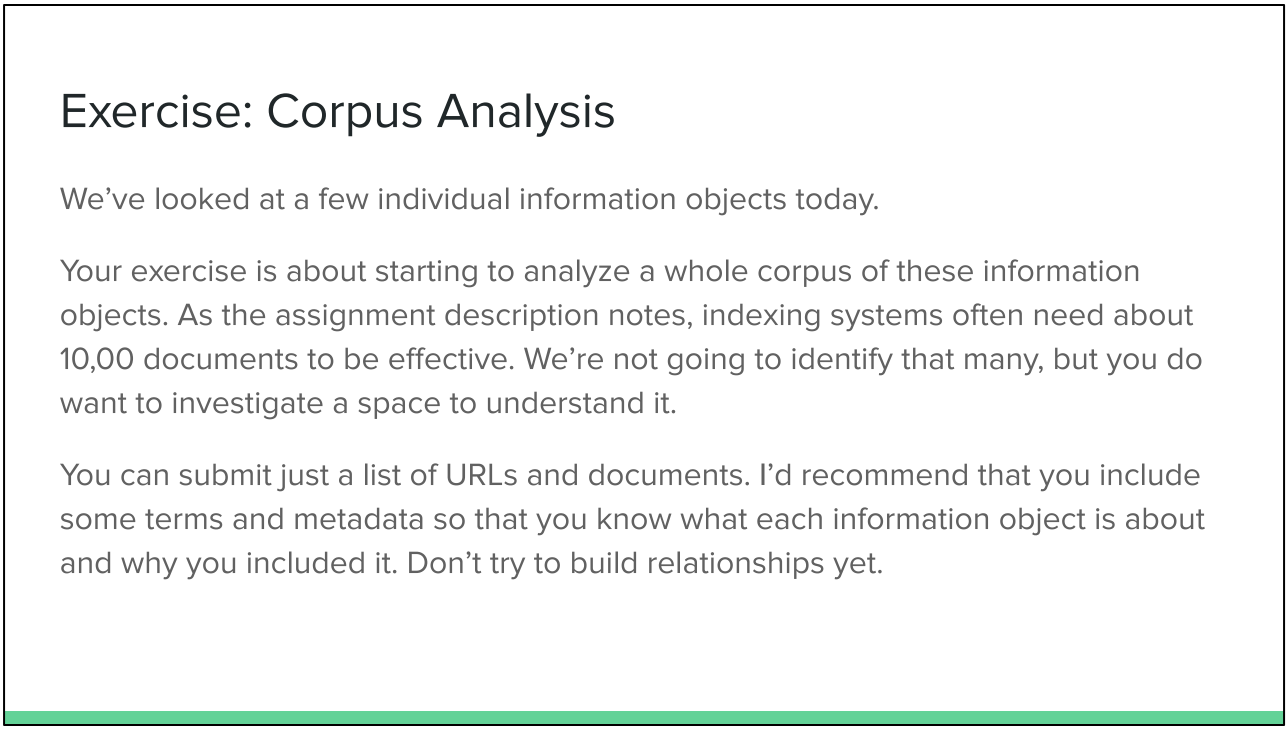
Consider precision and recall:

- Pre-coordinate indexing gives you high precision, but lower recall.
- Post-coordinate indexing gives you high recall, but lower precision.

Post-coordinate: Each term in a controlled vocabulary should represent a **single** idea. You can combine these ideas later either via a user doing a search or in an ontology with relationships.

Advice from Accidental Taxonomist, use pre-coordination when:

- Post-coordination isn’t working for retrieval.
- Subject area is focused and deep and you have specialized vocabulary.
- There’s a lot of content specifically about the pre-coordinated term.
- Highly varied document types, which makes post-coordination less likely to work.
- You’re creating a strictly hierarchical taxonomy, where it is difficult or impossible to combine terms as you might with an ontology or facets.



# Exercise: Corpus Analysis

We’ve looked at a few individual information objects today.

Your exercise is about starting to analyze a whole corpus of these information objects. As the assignment description notes, indexing systems often need about 10,00 documents to be effective. We’re not going to identify that many, but you do want to investigate a space to understand it.

You can submit just a list of URLs and documents. I’d recommend that you include some terms and metadata so that you know what each information object is about and why you included it. Don’t try to build relationships yet.