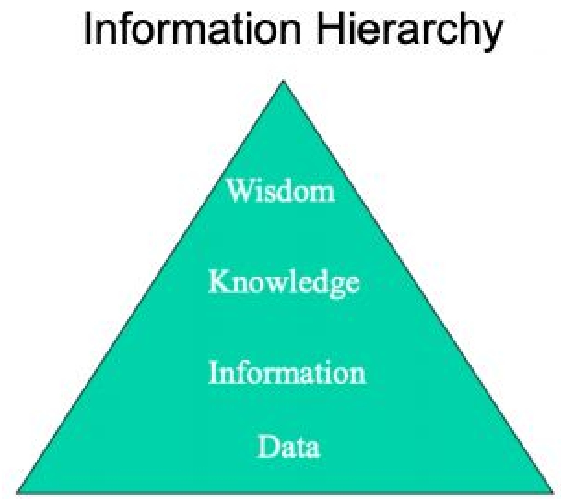
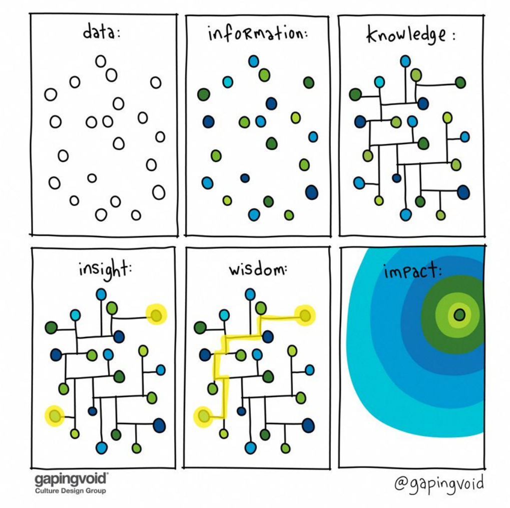
Information vs. Data vs. Knowledge vs. Wisdom

- What are the meanings of these terms?





From Lecture (09/01/2020)

1. **Data:** The raw material of information
   Elementary observations about properties of objects, events, and their environment.
2. **Information:** Data organized and presented by someone
   Data is aggregated, processed, analyzed, formatted, and organized to add meaning and context, and used to answer questions
3. **Knowledge:** Information read, heard or seen and understood
4. **Wisdom:** Distilled and integrated knowledge and understanding

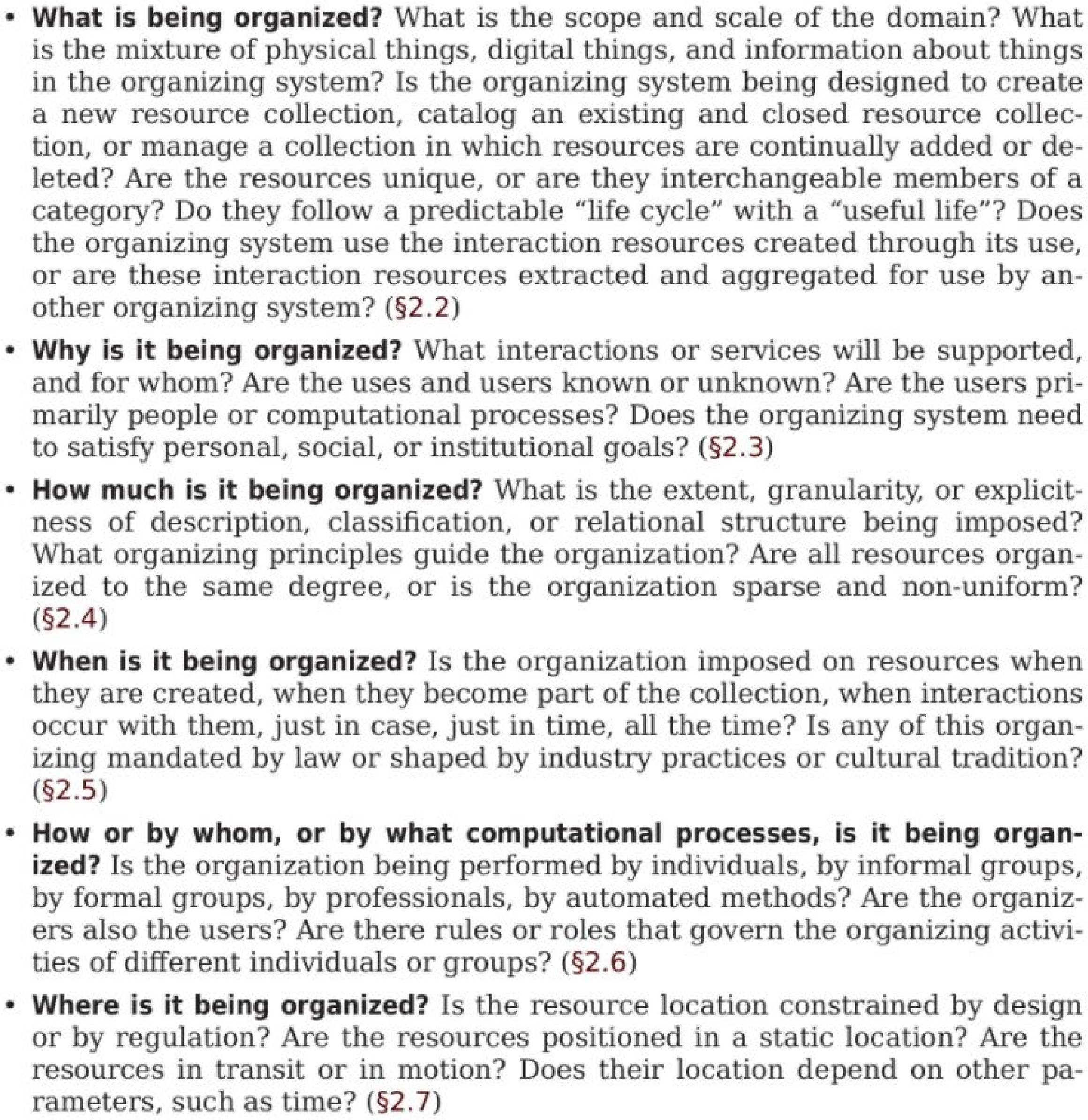
- How do they compare and contrast?
  - Consider the difference between asking a train conductor “Can I have some information about the train schedule?” vs. “Can I have some data about the train schedule?”
  - TDO p. 29
    - Data is transformed into Information, which is transformed into Knowledge, which is then transformed into Wisdom.

Design Decisions in Organizing Systems

- What, Why, How Much, When, Who/How
  - *Be able to analyze a collection in these terms*

- A set of resources is transformed by an organizing system when the resources are described or arranged to enable interactions with them.
- Resource: what is being organized
- TDO p.62 Design Decisions in Organizing Systems
- Lecture 5: Resource Selection
- **What** - selected resources
    - The same "thing" could be treated differently in different contexts
    - What's being organized? What's the scope and scale of the domain?
    - Is the organizing system being designed to create a new resource collection, catalog an existing, or manage one?
- **Why** - intended uses and intended users
    - Define overall goals, like maximizing efficiency, creating additional value with interactions, or influencing choices
    - Different organizations will have different purposes
        - Memory institutions: resource preservation
        - Management information systems: business needs
        - Government: legal requirements, transparency
    - The more we know about a resource + the more users → more use cases → potential for more diverse organizing systems and principles
        - Potential to lead to inequity
    - Why is it being organized? Are the users primarily people or computational processes?
- **How Much** - granularity of Annotations
    - Must consider detail applied to each resource, amount of organization into classes, and the extent of available interactions
    - What is the extent, granularity, or explicitness of description, classification, or relational structure being imposed?
    - What organizing principles guide the organization?
    - Are all resources organized to the same degree?
- **When** - organizing before search, or on demand. On the way in or on the way out?
    - Systems can change a lot over time, so adaptability is important
    - Is the organization imposed on resources when they are created, when they become part of the collection, when interactions occur with them, just in case, just in time, all the time?
    - "on the way in" = when they are created or made part of a collection
    - "on the way out" = organization imposed when an interaction with resources takes place
    - Google's "on the way" in selection and ranking algorithms for analyzing search results allow them to determine "on the way out" information about the user's search history and current context in the fraction of a second after the user submits a query
- **How/Who -** authors, automation, crowd, professionals, combination, etc.

        - Is the organization being performed by individuals, by informal groups, by formal groups, by professionals, by automated methods?
    - **Where -**
        - Is the resource location constrained by design or by regulation? Does their location depend on other parameters, such as time?



- **What is being organized?** What is the scope and scale of the domain? What is the mixture of physical things, digital things, and information about things in the organizing system? Is the organizing system being designed to create a new resource collection, catalog an existing and closed resource collection, or manage a collection in which resources are continually added or deleted? Are the resources unique, or are they interchangeable members of a category? Do they follow a predictable "life cycle" with a "useful life"? Does the organizing system use the interaction resources created through its use, or are these interaction resources extracted and aggregated for use by another organizing system? (§2.2)
- **Why is it being organized?** What interactions or services will be supported, and for whom? Are the uses and users known or unknown? Are the users primarily people or computational processes? Does the organizing system need to satisfy personal, social, or institutional goals? (§2.3)
- **How much is it being organized?** What is the extent, granularity, or explicitness of description, classification, or relational structure being imposed? What organizing principles guide the organization? Are all resources organized to the same degree, or is the organization sparse and non-uniform? (§2.4)
- **When is it being organized?** Is the organization imposed on resources when they are created, when they become part of the collection, when interactions occur with them, just in case, just in time, all the time? Is any of this organizing mandated by law or shaped by industry practices or cultural tradition? (§2.5)
- **How or by whom, or by what computational processes, is it being organized?** Is the organization being performed by individuals, by informal groups, by formal groups, by professionals, by automated methods? Are the organizers also the users? Are there rules or roles that govern the organizing activities of different individuals or groups? (§2.6)
- **Where is it being organized?** Is the resource location constrained by design or by regulation? Are the resources positioned in a static location? Are the resources in transit or in motion? Does their location depend on other parameters, such as time? (§2.7)

- What are the important choices and tradeoffs in designing organizing systems?
    - **Form vs. Function** - which should be the first step?
    - TDO p. 34

- **On the way in vs. On the way ou**t - should more effort be put into organizing the data to be easily retrieved, or should more effort be required to retrieve the data?
  - “The effectiveness of a system for accessing information is a direct function of the intelligence put into organizing it” - TDO Chapter 1
- **Intentional arrangement vs. naturally occurring patterns -** the emphasis of human intentionality and action (which can be both top-down or bottom-up)
  - For example, the stars weren’t intentionally arranged, but constellations were!
- Design decisions evoke a wide range of technical and social concerns
- Every organizing system is biased by the perspectives and experiences of the people who created it.

Information Architecture (IA)

See: TDO Page 109
(see lecture from class 3)

- What is it? What is it *not*?

  a. **Information Architecture:** the structural design of shared information environments; the way we organize parts of something to make it understandable; helps us make sense of the world around us

  b. From TDO: A specialized approach for designing the information models and their systematic manifestations in user experiences on websites and in other information-intensive organizing systems

  c. ○ is not algorithmic, needs judgement and practice

  d. ○ is not isolated to one field -> interdisciplinary nature

- What is the role of information organization in IA?

  a. Information organization depends on the users’ needs and business priorities. IA allows the user to reach the content easily without much effort

  b. IA can encompass everything including the organization of pages, overall site structure, organization of information, etc.

- How does content relate to structure and both relate to presentation? (Lecture 10)

  a. Content - “what does it mean”,

  b. structure - “how it is organized or assembled”,

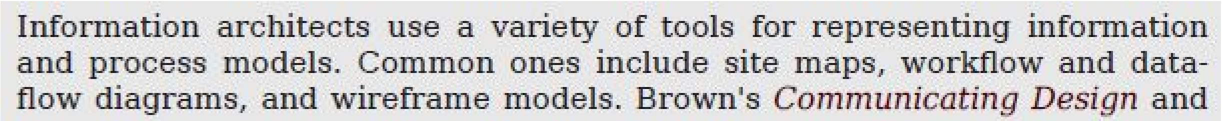
  c. presentation - “how it looks/how it is displayed”

- What tools do we use to do IA, and how do we use them

  a. *For example, card sorting, grounded coding*

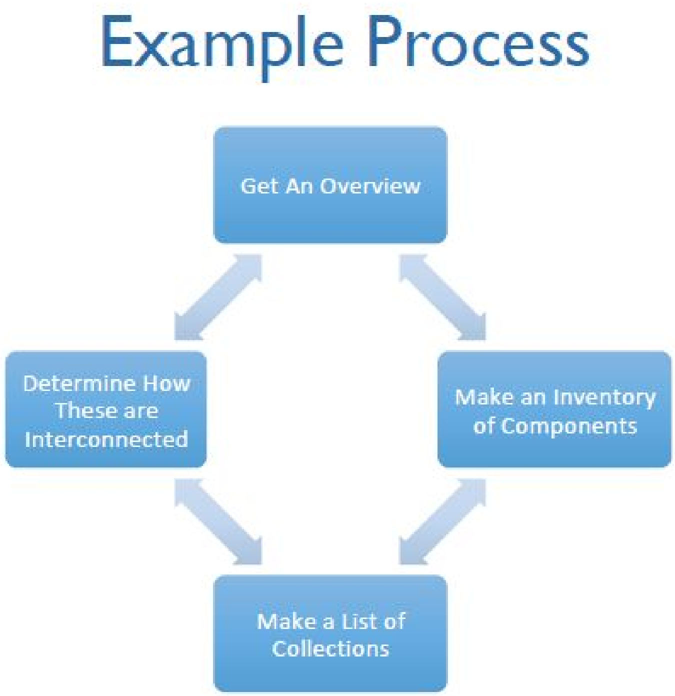
    - **Card sorting:** a way to uncover people's underlying assumptions about how people organize concepts (Lecture slides 3)
        - **Open Sort:** first given set of terms, then group them, then name groups/create categories
        - **Closed Sort:** first given pre-defined categories as well as unsorted term, then sort terms into categories; this could potentially lead to small or even empty categories
    - **Grounded coding:** method for assigning categories to qualitative data to make sense of it (Lecture slide 9)
        - Iterative process, with built consensus around coders
        - Use Cohen's Kappa to measure inter-annotator agreement
    - **Treesort/Tree-test:** a defined tree that we can use to test the success of our IA (Lecture slide 7)
        - Kinda like the optimalSort thing we did in class, with the bird??
    - http://www.howtomakesenseofanymess.com/

b. TDO pg. 111



Information architects use a variety of tools for representing information and process models. Common ones include site maps, workflow and data-flow diagrams, and wireframe models. Brown's *Communicating Design* and

- Be able to analyze the information architecture of a website.



Example Process

Get An Overview
Make an Inventory of Components
Make a List of Collections
Determine How These are Interconnected

Resources

- Properties: what are the individual resources (in a collection)? What is their granularity? How do we identify them? Which ones are identical? (Lecture Slides 5)
    - **Resource:** anything of value that can support goal-oriented activity; an entity that is the subject of organization (TDO p.65)
    - TDO p. 162) Resources are what we organize. A group of resources can be treated as a collection.
    - Selection is shaped by the domain and then by the scope (so their granularity) of the organizing system, which can be analyzed through six interrelated aspects:
        1. The number and nature of users
        2. The time span or lifetime over which the organizing system is expected to operate
        3. The size of the collection
        4. The expected changes to the collection
        5. The physical or technological environment in which the organizing system is situated or implemented
        6. The relationship of the organizing system to other ones that overlap with it in domain or scope
    - **Metadata:** resources associated with other resources.
    - We can use a resource's unique identifier (UID) to refer to it unambiguously
    - TDO p. 165~166
    - **Vocabulary Problem:** multiple names for the same resource; solve partially with a controlled vocabulary (TDO Chapter 4 page 165) and mapping so that everything points to what should be considered the same thing

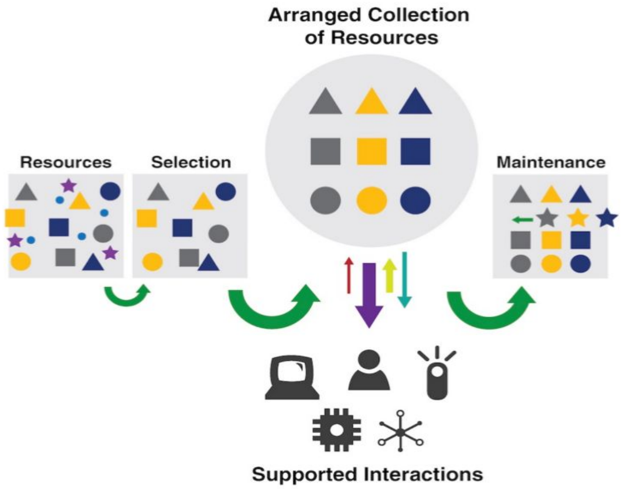
- Selection: what is the process? What are the principles?
    - **Selection:** when resources are identified, evaluated, and added; an intentional process; methods and criteria vary across domains and can be subjective or objective (TDO p.92)
    - Selection must be an intentional process because by definition, an organizing system contains resources whose selection and arrangement was determined by human or computational agents, not by natural processes.
        - By going **upstream,** we can consider the source of the resource and understand why something is organized in a particular way (TDO p.96)
        - If we consider what is **downstream**, we account for how our resource might be used in the future. Analyzing any evidence or records of data's use or interactions. (TDO p. 97-98)
        - **Selection principles: (Lecture slides 5)**
            - Utility, usefulness, relevance
            - Comprehensiveness
            - Intrinsic Value
            - Scarcity or uniqueness

  - To support social goals
  - To establish a reputation or brand
- Selection methods can be subjective or objective
- Selection **is the process by which resources are identified, evaluated and added to a collection**
- Recognition over recall, it's usually easier for a person to recognize something by looking for it than it is to think up how to describe that thing.

- Provenance: what is this? Why does it matter?
  - Pg. 198, 203 TDO, Lecture slides 2 (finding image source)
  - **Provenance:** considering *where* the resource came from, *what* has happened to it since, *who, what, where, when, why,* and *how?*
  - "Provenance is the history of the ownership of a collection or the resources in it, where they have been and who has possessed them."
  - It is important to give attribution to the resources used and prevent anything bad from happening to them
  - Understanding provenance helps you evaluate whether a resource has maintained its quality over time

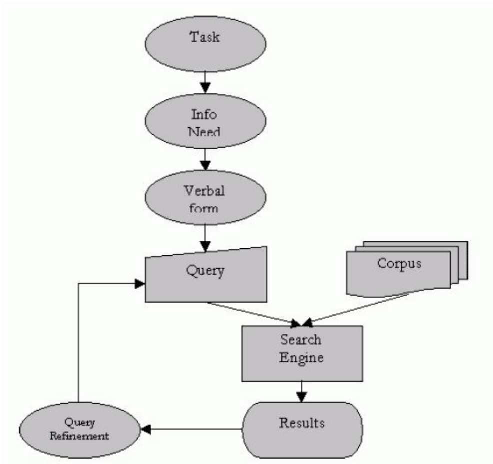
- Maintenance: why should we care about this? (TDO p.90, 133)
  - What governance policies and procedures are needed to satisfy retention, compliance, security, and privacy requirements?
  - Maintenance activities are usually described as preservation or curation
  - Often described as deletion, purging, data cleansing, governance, or compliance
  - Important because: resources change and are updated overtime, therefore the collections should be updated as time goes on

Figure 3.1. Four Activities in all Organizing Systems.

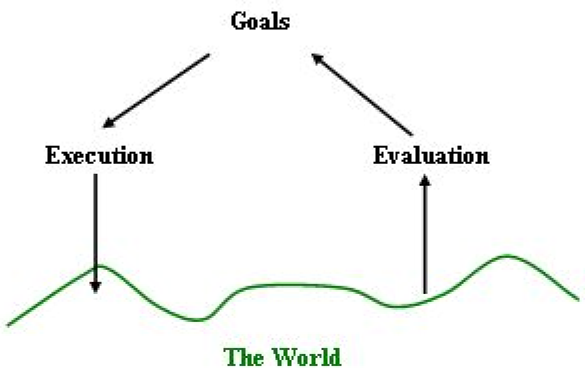


Four activities take place in all organizing systems: selection of resources for a collection; intentional organization of the resources; design and implementation of interactions with individual resources or with the collection, and; maintenance of the resources and the interactions over time.

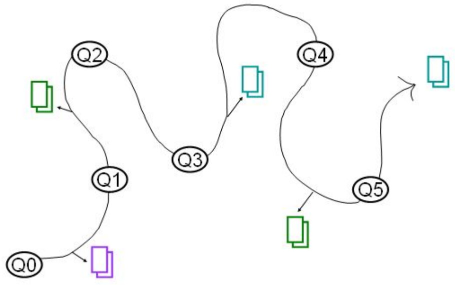
Information Seeking and Search Interfaces

- What are the primary models of information seeking?
  - **The Standard Model - Sutcliffe & Ennis 1998 (SUI Ch.3)**
    - Identifying an information need → activities of query specification → examination of retrieval results → reformulation of the query → repeat cycle until result is found
    - Has 4 main activities
      - Problem identification
      - Articulation of Information needs
      - Query formulation
      - Results Evaluation
    - Standard Web search engines support query specification, examination of retrieval results, and query reformulation
    - User's information need is static and the information seeking process is refining a query until all documents needed are retrieved
    - 

  - **The Cognitive Models**
    - A person must first have a basic idea of what they want (i.e the goal to be achieved) → the person then uses a *mental model* to decide on an action that affects someone or something to achieve their goal
      - *Mental model* meaning one's understanding of the system or interface



        -
        - Execution ~ doing, Evaluation ~ checking of the result
        - After taking an action, a person must assess what kind of change occurred and whether or not the action achieved the intended goal
        - This also suggests that the less knowledge a person has about their task, the less they will be able to successfully formulate goals and assess results.
        - Recognizing a need for information is analogous to formulating and becoming conscious of a goal.
        - Query reformulation is needed if the gulf between the goal and the state of the world is too large.
    - **The Dynamic/Berry-Picking Model**
        - Searchers' information needs change as they interact with the search system, and/or goals change in priority
        - Searchers learn about the topic as they scan retrieval results and formulate new sub questions as previhous questions are answered
        - **Berry Picking Model**:
            - In the process of reading and learning from the info encountered throughout the search, the searcher's info needs are constantly shifting
            - Searcher's info needs are not satisfied by a single set of retrieved documents, but instead by a series of selections.

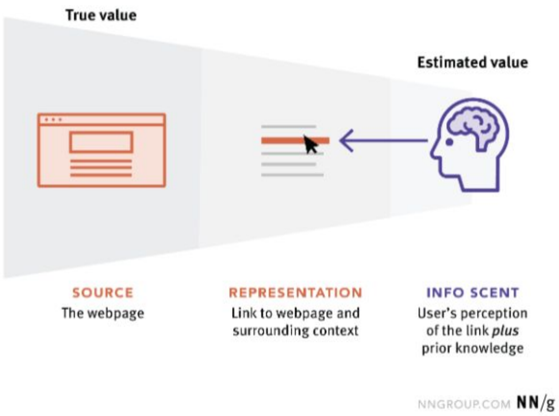


            -

    - Six Stages of Information Seeking (SUI 3.4: INFORMATION SEEKING IN STAGES):
        - Initiation: recognize need for info; feeling of awareness
        - Selection: select topic or approach; feeling of optimism
        - Exploration: investigate information to extend understanding; feeling of confusion + doubt
            - More connected with general browsing
        - Formulation: emergence of focused perspective; resolve the previous confusion + doubt
            - (you know your topic but you have to come up with a very specific question)
            - Forming an opinion of general browsing
        - Collection: gather info around focused topic; define, extending, and supporting your focus
        - Presentation: info search becomes redundant, so there is a general feeling of relief

- What are search strategies?
    - Search strategies are made up of sequences of search tactics. Strategies refer to combinations of tactics used in order to accomplish info access tasks
    - **The Sensemaking Model:** process of searching for a representation and encoding data in that representation to answer task-specific questions; the iterative process of forming a conceptual representation from a large volume of info
    - Search tactics in 4 categories:
        - **Term Tactics**
            - Tactics for adjusting words and phrases within the current query
                - refining words in keyword searched. Influenced by the results that you get.
        - **Information structure tactics**
            - Techniques for moving through information or link structures to find sources or information within sources
        - **Query Reformulation Tactics**
            - Narrowing a given query specification(Getting your query closer to your information need) by using more specific terms
            - Gaining more control over the structure of the query by using Boolean operators
        - **Monitoring Tactics**
            - Keeping track of a situation as it unfolds
            - Ex. comparing the current state with the original goal & recognizing patterns across common strategies

- How is navigation integrated with search user interfaces? (SUI Ch.8)
    - **Navigation/Browsing:** selecting links or categories that produce pre-defined groups of information items
    - Browsing can also refer to casual, undirected exploration of navigation structures
    - Search queries usually produce new collections of info that haven't been gathered together before
    - Best practices in information architecture emphasize the use of systematic principles or design patterns for organizing resources and interactions in user interfaces
    - Some design conventions have become patterns - large text signifies bigger category, bold says pay attention
    - Design patterns reflect and reinforce the user's past experiences with content and interface, this familiarity reduced cognitive complexity
    - 'A weakness of many Website's' navigation structure is that the Web sites do not integrate their browsing with their searching functions - the google system uses faceted categories only for narrowing search results; only one category's information can be seen at any time.'

- How do choices in categorization and resource naming intersect with information scent? (SUI 3.5.4)
    - **Information Scent:** cues that provide searchers with concise info about content that's not immediately noticeable
    - Search results listings must provide the user with clues about which results to click
    - Spool suggests that by showing users informative hints about what kind of info can be found "one hop away," we can operationalize information scent
    - 
True value
Estimated value
SOURCE
The webpage
REPRESENTATION
Link to webpage and surrounding context
INFO SCENT
User's perception of the link *plus* prior knowledge
NNGROUP.COM NN/g

Categorizing Information

- What are category systems good for? (TDO p.325)
    - **Categories:** sets or groups of resources that are treated the same; division and naming of the physical/experienced world
        - As opposed to *classification*, which are the principles used to define and fill categories
        - Allows us to cognitively interact with novel things that belong in categories of things we've seen before.
        - Enable us to relate things to each other in terms of similarity and dissimilarity and are involved whenever we perceive, communicate, analyze, predict, or classify.

- How do Category Systems fail?
    - Don't capture relational links well
    - Unintuitive information
    - Content and categories don't match
    - Too much hierarchy (too many things to click through)
    - Can be biased

- What are good strategies for organizing information into categories? (TDO 7.2.1~7.2.5)
    - **Cultural -** based on everyday experience of the world; wide variation across cultures
    - **Individual** - satisfy ad-hoc requirements that arise from an individual's unique experiences → short-lived; draw from cultural categories, but with the addition of the individual's imagination
    - **Institutional -** systematically created by domains
    - Color example: cultural ~ association of anger with red, individual ~ color blindness + perception, institutional ~ HEX code or primary colors
    - **Computational** categories are created for information retrieval, predictive analyses, and other applications where information scale or speed requirements are critical. We can also use computational processes (ie. ML)

- What is difficult about assigning categories to information items? (TDO 7.4)
    - Abstraction vs. Granularity - how specific should we get?
    - Some categories are more "natural" than others
    - Recall vs. Precision

- Difference between categorization and classification (TDO 8.1)
    - Classification is the "rules" for the category

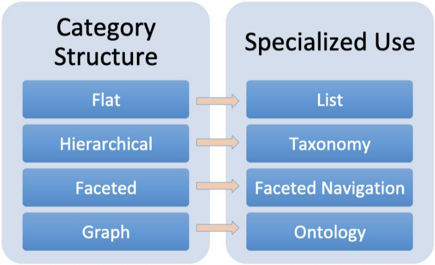
    - Giving structure and system to the categories
    - **classification** is the act of forming into a class or classes; a distribution into groups, as classes, orders, families, etc, according to some common relations or attributes
    - while **categorization** is a group of things arranged by **category**
    - Categorization: describing resources classes and types
    - Classification: assigning resources to categories
- What are principles for classification? (TDO 8.2)
    - *Be able to describe how they are applied to an example classification*
    - Lecture slides 6
    - **Enumeration:** assigning sequential numbers; good for unambiguous membership, but bad for inflexibility and impracticality with large sizes (ie. ISO) (TDO p.337)
    - **Property-Based Categorization**: intentional definition; as more and more properties are applied, it becomes harder and harder for humans to cognitively keep up
        - **Single Properties:** one property or rule (TDO p.338)
            - **Intrinsic static property:** inherent in a resource, never-changing
                - Ex. The number of times a word appears in a document. The number of times a letter exists in a word
            - **Extrinsic static property:** arbitrarily assigned, never-changing
                - Ex. Social Security number, student ID number
        - **Multiple Properties:** >1 property to define scope (TDO p.340)
            - **Multilevel/Hierarchical:** use some sequence of resource properties (ie. separating shirts from pants, then short-sleeves from long-sleeves)
            - **Different properties for different subsets of resources:** each domain is conceptually adjacent, but is based on resource properties within that domain (ie. computer folders)
            - **Necessary & sufficient:** creating categories based on features that are only necessary and sufficient for identification (ie. prime numbers)
    - **Probabilistic Co-Occurrence of Properties:** categorization depends on accumulation of evidence from properties that are characteristic of the category (ie. birds)
        - Consequences for categories when their categories have a probabilistic distribution:
            - typicality/centrality: makes some members of the category better examples than others
            - family resemblance: the sharing of some but not all properties
            - unfixed boundaries: the category can be stretched and new members assigned as long as they resemble incumbent members

            - Ex: A spam classifier uses the probabilities of each word in a message in spam and non-spam contexts to calculate an overall likelihood that the message is spam
    - **Similarity:** measure of resemblance between two entities that share some features, but aren't necessarily identical
        - **Feature-based Similarity:** based on not only features shared, but also features unshared
        - **Geometric Similarity:** vectors represent items in multi-dimensional space, where property values determine where each point is
            - Green circle, red square example
                - the "redness" and the "greenness" of both objects are weighted differently
                - the shape of the circle and the shape of the square also gets weighted differently
                - Measure distance between two things. Where these things are on the x,y axis is determined by their features
        - **Transformational Model:** similarity between two things is inversely proportional to the complexity needed to turn one thing into the other
            - Ex. how many times can you change the word book to make bob
            - "book" and "bob" book —> boo —> bob
    - **Goal-Derived:** organizing resources to satisfy a goal; ad-hoc categorization (ie. *things I'd save from my house during a fire*)
    - **Theory-Based:** fits a theory or story that makes categorization sensible
    - Some choices are motivated, some more arbitrary
    - Consider purpose, usage, scope, lifetime, extensibility, pre-existing category systems
    - Be consistent once classification decisions have been made

Types of Categorization Systems

- What are: Flat, Hierarchical, and Faceted structures? How are they defined? (Lecture slides 6, 7
    - **Flat:** can be ordered or unordered
        - Alphabetized, by importance, cost, frequency, etc.
    - **Hierarchical:** one entity type contains or is comprised of other entity types; "is a," "is a part of," "is in," or "is a type of" relationships
        - Computer files, table of contents
        - **Taxonomy:** "is a" and "is a type of" relations; inclusion based organization
            - Biological classification (kingdom, phylum, class…)
    - **Faceted:** A set of categories to identify different aspects or features; resources are identified by multiple categories; good for searching and browsing

        - Cleanly separated facets allow for powerful navigation
        - Unfortunately, facets fail to show relational information
    - **Tags:** user-driven categorization; one item can have multiple tags
    - **Ontology** (TDO p.286): relationally-linked objects to support thematic relations; difficult to show and convey how to navigate (ie. IMDB); not really a form of categorization?
        - Uses syntax like married_to(x,y)
        - One can do inference over these relations
        - Relations can be inclusion type, attribution type, or possession type



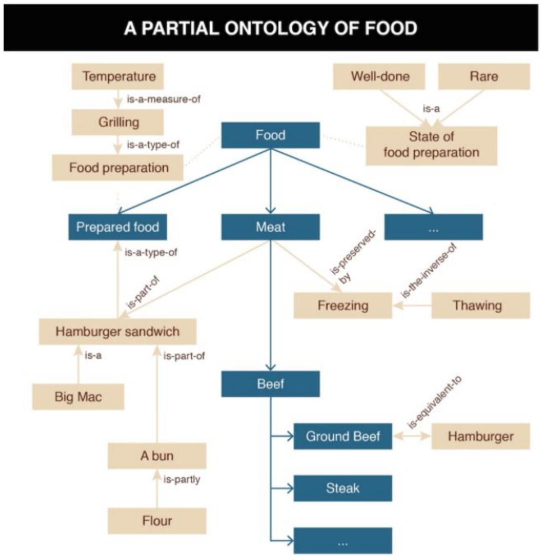
- What is Taxonomic vs Thematic categorization?
    - **Taxonomic:** grouping by type; based on shared or similar features; "is-a" relations; favored by young adults
        - Egg + Toast are in a group
    - **Thematic:** grouping by experience; based on co-occurrence in context; favored by children and older adults
        - Toaster + Toast are in a group

- What is Faceted vs Hierarchical categorization?
    - Faceted categorization can assign multiple categories to a single item based on different features, while hierarchical categorizations are encapsulated, "Russian doll" categories
    - Each faceted category *can* be hierarchical
        - **Hierarchical Faceted Metadata:** each faceted category is further organized by hierarchical categories

- What is Faceted Categories vs Tags?
    - Tags are usually flat, while faceted categories are often hierarchical
    - Tags are user-driven, and as a result, any term is okay. On the other hand, facets are controlled by the owner of the collection and thus have a controlled vocabulary.

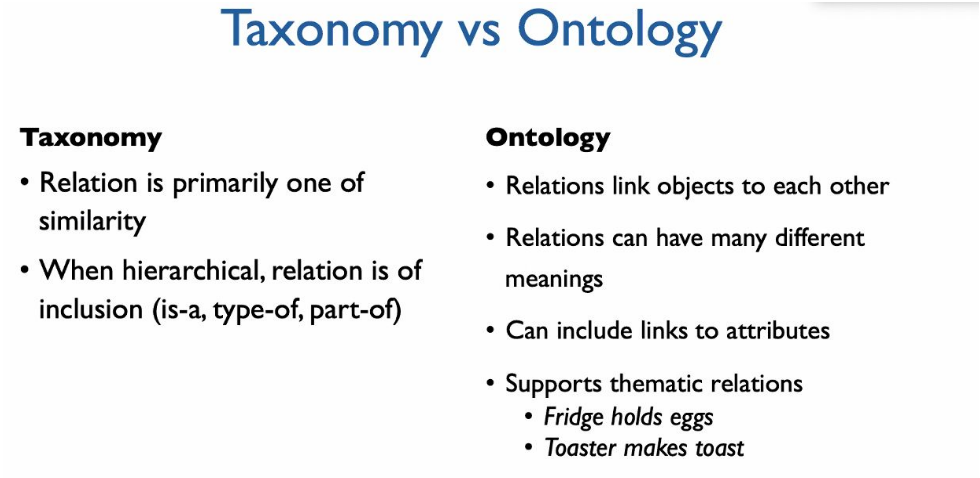
- What is Taxonomy vs Ontology?
    - Ontologies are based on relation links between objects, while taxonomies are based on inclusion relationships
        - These relations can have different meanings
        - Is-a
    - Thus, taxonomies support taxonomic relations, while ontologies support both thematic relationships and taxonomic
        - Ex: picture shown with meat/animal/grill/buns/etx. Ontology of a hamburger lol.
    - Be able to identify the category system being used in an information organization

Figure 6.2. A Partial Ontology of Food.



A partial ontology of food overlays the taxonomy of food with statements that make assertions about categories, instances, and relationships in the food domain. Example statements might be that "Grilling is a type of food preparation," that"Meat is preserved by freezing," and that "Hamburger is equivalent to ground beef."

- Be able to create category systems given information resources.

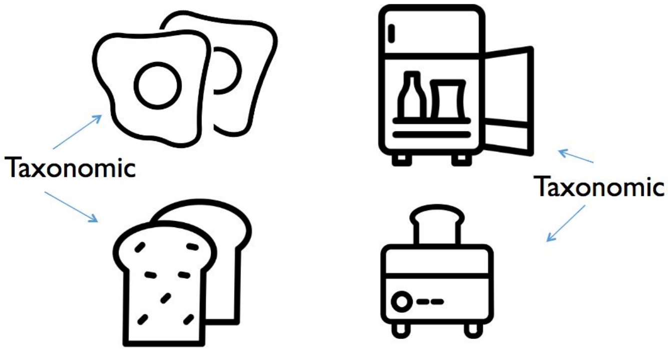
    - 

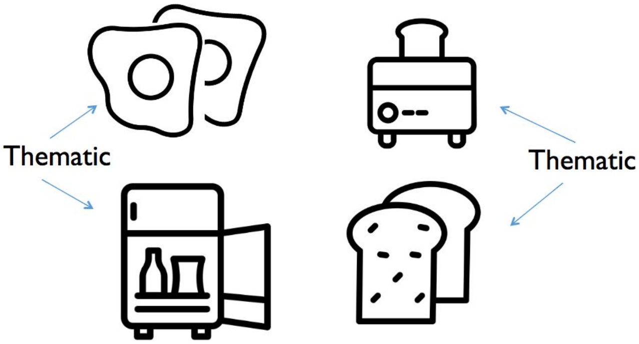
      ## Taxonomy vs Ontology

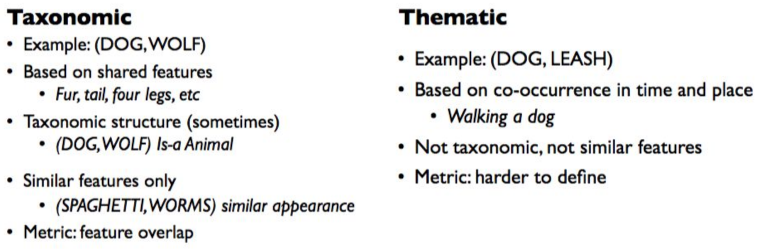
      Taxonomy
      - Relation is primarily one of similarity
      - When hierarchical, relation is of inclusion (is-a, type-of, part-of)

      Ontology
      - Relations link objects to each other
      - Relations can have many different meanings
      - Can include links to attributes
      - Supports thematic relations
          - *Fridge holds eggs*
          - *Toaster makes toast*

      Egg toaster fridge toast example

    - 



Taxonomic
- Example: (DOG, WOLF)
- Based on shared features
    - *Fur, tail, four legs, etc*
- Taxonomic structure (sometimes)
    - *(DOG, WOLF) Is-a Animal*
- Similar features only
    - *(SPAGHETTI, WORMS) similar appearance*
- Metric: feature overlap

Thematic
- Example: (DOG, LEASH)
- Based on co-occurrence in time and place
    - *Walking a dog*
- Not taxonomic, not similar features
- Metric: harder to define

    - Having to organize our comfort food

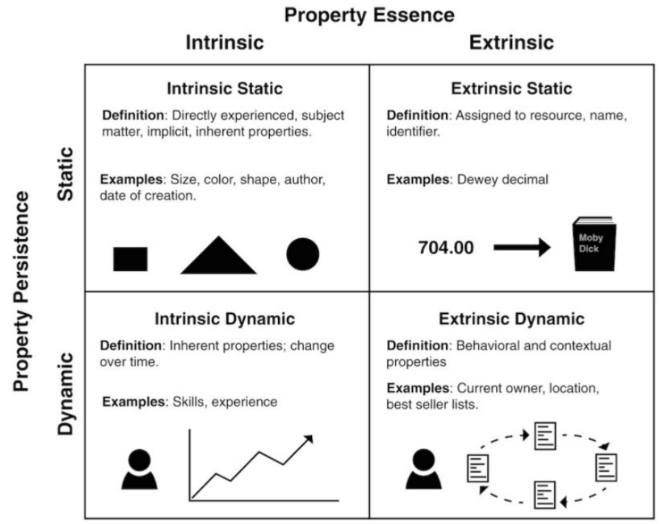
Resource Descriptors

- What are Controlled Vocabulary, Authority Control?
    - **Controlled Vocabulary** (TDO p.249)**:** a set of predetermined terms that describe specific concepts; a fixed or closed set of terms in a domain with precise definitions used instead of vocab that people would otherwise use
        - Using controlled vocabulary terms in your search strategy allows you to locate citations no matter what term an author does or doesn't use.
        - It also helps account for spelling variations and acronyms
        - Aims to be standardized and consistent to aid in search
        - Dictionaries, code lists, etc.
    - **Authority Control:** process that organizes bibliographic info in library catalogs by using a single, distinct spelling of a name (access point or heading) or a subject for each topic (TDO p.195)

- What is the process of Resource Description?
    - *Be able to complete any part of the resource description process.*
    - Resource Descriptions are used by people and computational agents and are designed to help people orient themselves
    - Process of describing resources that has 7 steps.
    - 7 Steps (From: 5.3. The process of describing resources p.227):
        - Determine the scope and focus of resources
            - Granularity or Abstraction
        - Determine the purpose
            - Resource Domain or Context of Descriptions
        - Identify resource properties by studying resources
            - Intrinsic vs Extrinsic Properties or Dynamic vs Static Properties
            - Intrinsic properties: those that are inherent no matter what context it exists in (no matter if helen is categorized as a student/human/woman, she's a helen??)

        - Extrinsic properties: SSN, CalNetID

Figure 5.4. Property Essence x Persistence: Four Categories of Properties.



		Property Essence	
		Intrinsic	Extrinsic
Property Persistence	Static	**Intrinsic Static** **Definition**: Directly experienced, subject matter, implicit, inherent properties. **Examples**: Size, color, shape, author, date of creation.	**Extrinsic Static** **Definition**: Assigned to resource, name, identifier. **Examples**: Dewey decimal 704.00 → Moby Dick
	Dynamic	**Intrinsic Dynamic** **Definition**: Inherent properties; change over time. **Examples**: Skills, experience	**Extrinsic Dynamic** **Definition**: Behavioral and contextual properties **Examples**: Current owner, location, best seller lists.

The distinctions of property persistence and property essence combine to distinguish four categories of properties: intrinsic static, extrinsic static, intrinsic dynamic, and extrinsic dynamic properties.

    - Design the description vocabulary
        - Make use of existing standards
    - Design the description form and implementation
        - Notation or Syntax or Structure
    - Create the descriptions
        - By people or by Algorithms
    - Evaluate the descriptions

Lexical Relations and Naming

- What is the Vocabulary Problem? (TDO p.166)
    - When a resource or concept has more than one name or identifier
    - People use different words to describe ideas
    - An example used in class was the lady with the banana moustache → different people had different ways of describing the photo
    - People think their own vocabulary is what is “intuitive” or “natural,” which leads to surprise when this arises

- Why is it difficult to choose names for labels/categories that are universally understandable? (TDO p.288)
    - **Name:** a label for something or category that distinguishes it from another
        - **Identifier:** a special name assigned in a controlled way and governed by rules

- Lexical gaps differ from culture to culture, in that some languages may lack a word for a concept that another language has
    - For example, Chinese has a specific word for older sister

What is semantic vs lexical? (Lecture 9)

- *Lexical* refers to the specific word, while *semantic* refers to the concept
- Lexical ~ how it is expressed, semantic ~ the meaning

What are the main types of lexical relations? (Lecture 9)

- *Be able to come up with and identify examples*
- **Hyponymy and Hyperonymy:** hyponym is the more specific class, and the more general class it belongs to is the hypernym
    - “Robin” vs. “bird”
    - “Is a” hierarchy → taxonomies (inclusion relationships (categorization))
- **Meronymy and Holonymy:** part(meronymy) - whole(Holonymy) relations
    - The engine is-part-of the car
    - The book is-part-of the library
    - The slice is-part-of the pie
    - “Wire” vs. “trap”
    - Not for taxonomies, for ontologies
- **Metonymy:** an entity is described as something that is contained in or a part of it
    - Using “Wall Street” to refer to the economy
    - Saying the White House to refer to the President and his team
    - Hollywood to refer to film industry
    - substituting the name of an attribute or feature for the name of the thing itself.
- **Synonymy:** same semantic concept
- **Polysemy:** a word with several different meanings or senses in context
    - “Bank,” “Shipping container”
- **Antonymy:** opposite meaning

What are the types of structure for Resource Identifiers? (Lecture 8)

- *Be able to identify choices made for structuring resource identifiers, and describe motivations for them.*
- An identifier is ***unique*** if it refers to one and only one resource within a defined context
- An identifier is ***persistent*** if it resolves the referent indefinitely (or as long as it is necessary)
- An ***unstructured*** identifier has no inherent meaning based on its values (ie. student ID’s)
- A ***structured*** identifier has meaning, though it can become less meaningful over time (ie. website URL’s)

        - Highly structured identifiers provide more specificity (ie. addresses in Japan)
    - To what degree do resource identifiers have meaning vs. no meaning (e.g. discussion on class on whether phone numbers, car model numbers and SSN have meaning and to what degree)

Metadata and Metadata Descriptors

- What is metadata, and what is it for?
    - **Metadata:** a standardized resource description; often a functional substitute for the resource it describes when the resource itself can't be accessed.
    - Metadata is "data that provides information about other data".
        - Used for search and automation

- When you encounter a dataset, how should you inspect its metadata?
    - Kaggle (delayed flights) exercise in class
    - Look at columns one by one, for each column inspect possible values and perhaps even distribution of these values to make sense of labels

- What is a metadata descriptor? What is markup?
    - We can use query languages like XQuery and JSON with metadata descriptors to search for items.
    - TDO distinguishes between the "metadata" / "associated resource" versus the "primary resource" itself in the earlier chapters
    - TDO explicitly says that tags are "user-defined and could be anything"

    - **Metadata descriptor: <what's inside here> (heading in DTD for facetmap)**

        - <!ELEMENT Name (#PCDATA)>
        - Metadata: Updated Date
        - Metadata descriptor <updated_date>{Updated Date Value}</>
    - **Markup:** any language that uses tags to define elements in a document; usually human-readable

- What are standards, and how do they relate to a metadata description language?
    - **Standards**: a requirement which is intended to establish a common understanding of the meaning or semantics of the data

- **Metadata description language** (MDL), a common interchange **language** that describes requirements, design choices, and configuration information
  ○ Need conceptual modeling and analysis first → XML seeks to represent that and make it happen
  ○ **XML schema -** defines possible types of content in a doc and rules that govern structure and values of that content
  Must have common intention between producer and consumer of XML document (so can read and receive)
  we CAN format an XML schema to be a metadata descriptor, but we can't assume that every XML element fits our definition of metadata

- What are the core aspects of XML?
  - How does it differ from HTML? How is it similar? (Lecture 10)
    - HTML involves a fixed set of format-oriented tags that are meant to be processed by the human eye.
    - XML has a tag set decided on by the owner, and it is meant to be processed by the computer.
    - **XML** and **HTML** markup languages **are** related to each other where **HTML** is used for the data presentation whereas the main purpose of **XML** was to store and transfer the data.
    - **HTML** is a simple, predefined language while **XML** is the standard markup language to define other languages. **XML** document parsing is easy and fast.
    - XML Separation between info modeling and UI interface/graphic design
    - XML can realize document models suitable for implementation in applications

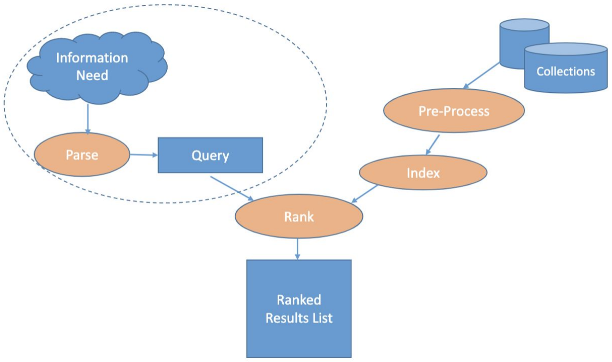
  - Be able to understand XML specifications, markup and DTDs
    - What does XML express?
      - Formatted data
      - Content and organization only
    - What does XML not express?
      - Not visual representation

    - **DTD/Schema:** Defines a document's possible types and the rules that structure/find value in that content; what the XML doc is validated against; includes entity declarations, element declarations, and attribute declarations
      - DTD = document type definition

        - A **markup** language is a computer language that uses tags to define elements within a document. It is human-readable, meaning **markup** files contain standard words, rather than typical programming syntax.

    - Be able to produce simple versions of these
        - **XML:** generic language for describing the content and structure of documents, as well as the markup that can be used for these documents
        - XML Big Ideas (Advantages and Cool things we like)
            - It's extensible - you can create a new set of tags for more specific contexts
            - Content and presentation are separate
            - Schemas define document types
            - Schemas enable XML documents to be validated
            - XML usually comes from non-XML information
            - Up translation -> convert to XML doc; Down translation -> convert to HTML, DB table, print doc

Information Retrieval



- What is the overall structure of how documents are processed for information retrieval?
    - Text Processing Steps:
        - Recognize document structure
        - Break into tokens
        - Stemming/morphological analysis
            - **Stemming:** converting different variations of a word to its stem (ie. operate, operates, operating → oper)
        - Store in inverted index

    - Why do we represent documents as vectors?
        - **Vector Representation:** documents are considered "bags of words;" each document is a vector in a multidimensional space, where the vector

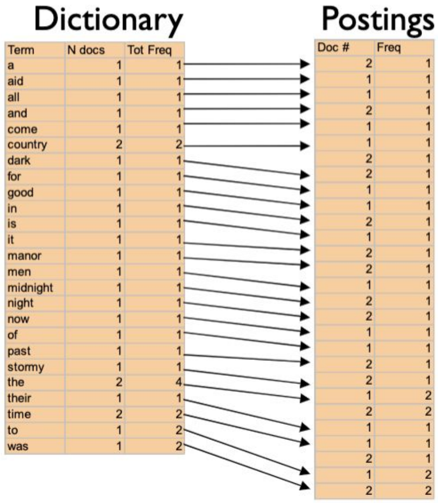
        is computed based on terms that exist in the entire *collection of documents* (some terms will have 0 weight because it doesn't exist in the document!)

        - If we represent a query as a vector as well, we can use distance between the query and each document to determine what to return
        - **TF x IDF:** a method for assigning weights to terms; used in vector space model to compare for similarity
            - **BM-25:** combined weight of term *i* in document *j* is

$$= \frac{CFW(i) * TF(i,j) * (K1 + 1)}{K1 * ((1 - b) + (b * NDL(j))) + TF(i,j)}$$

                -
                - Where CFW(i) is collection frequency weight log(#docs in collection) - log(#docs term *i* occurs in)
                - TF(i,j) is the frequency of term *i* in doc *j*
                - NDL(j) is the #terms in doc *j* divided by the average document length of all documents
                - K1 is a tuning constant for term frequency
                - b is a turning constant for document length

- Why do we build inverted indexes?
    - **Inverted index:** make a dictionary, where key:value pairs are "term, lengthOfDocArray, total frequency": Array of pair values (docID, frequency in docID) (IR p.7)
    - First parse through each document to extract terms and save as a dictionary of term:docID pairs. Then, perform intersection on them.
    - For each term, you get a list of docID, frequency of that term in the doc (optional), and the position of the term in the doc (optional)
    - Permit fast search for individual terms



Dictionary

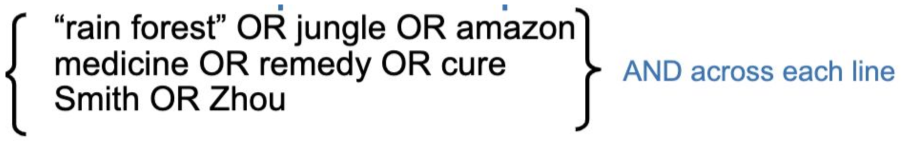
Term	N docs	Tot Freq
a	1	1
aid	1	1
all	1	1
and	1	1
come	1	1
country	2	2
dark	1	1
for	1	1
good	1	1
in	1	1
is	1	1
it	1	1
manor	1	1
men	1	1
midnight	1	1
night	1	1
now	1	1
of	1	1
past	1	1
stormy	1	1
the	2	4
their	1	1
time	2	2
to	1	2
was	1	2

Postings

Doc #	Freq
2	1
1	1
1	1
2	1
1	1
1	1
2	1
2	1
1	1
1	1
2	1
1	1
2	1
2	1
1	1
2	1
2	1
1	1
1	1
2	1
2	1
1	2
2	2
1	1
1	1
2	1
1	2
2	2

    -

- What are Boolean queries?
    - How do they work? How are they related to faceted navigation interfaces?
    - Consists of terms and their connectors/operators
        - AND, OR, NOT
        - Use parentheses to group things
    - Unfortunately, pure boolean search does not yield ranked results, so in practice, results are ranked chronologically, by order of total "hits," etc.
    - **Faceted Boolean Query:** breaking query into facets with a giant AND
        - The Wayfair example we did in class

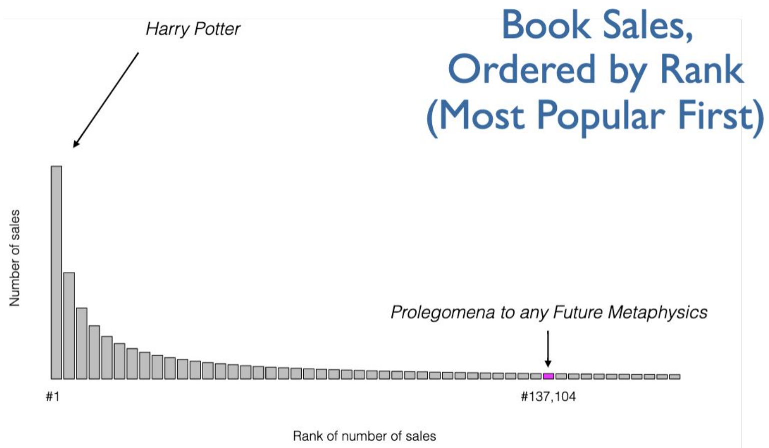


"rain forest" OR jungle OR amazon
medicine OR remedy OR cure
Smith OR Zhou

AND across each line

        -
        - Basically what filters are
    - We can also use aspects like proximity, nearness, phrases, and phrase variants in our search

- What is Zipf's law? Why does it matter for weighting and ranking terms for information retrieval?
    - **Zipf's Law:** there is a relationship between frequency of an event and the rank of that frequency among all events; therefore, the distribution of terms within a document is not uniformly or normally distributed
        - Count the number of times a term occurs in a document
        - Rank these terms by how often they occur



        -
            - A few elements occur very frequently, while a lot of elements occur infrequently
            - The most interesting words for information retrieval fall somewhere in the middle

        - the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. For example, the word "the" is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences. the second-place word "of" accounts for slightly over 3.5% of words.
    - The product of the frequency and rank is approximately constant
    - Helps us identify “stop words”
        - Ex. “a”, “the” words used frequently in the english language but are not descriptive of what you’re looking for

Relevant readings

http://www.howtomakesenseofanymess.com/

https://medium.com/@b.terryjack/nlp-everything-about-word-embeddings-9ea21f51ccfe