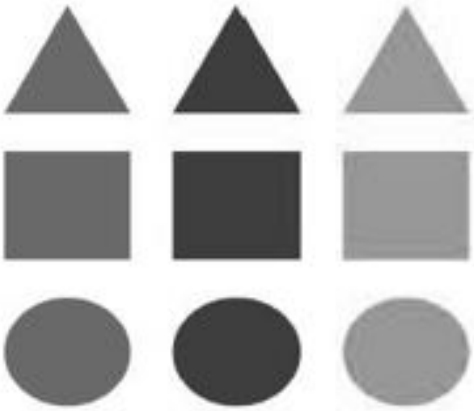
THE DISCIPLINE OF ORGANIZING

CORE CONCEPTS EDITION



Edited by **ROBERT J. GLUSHKO**

Chapter 8

Classification: Assigning Resources to Categories

Robert J. Glushko
Jess Hemerly
Vivien Petras
Michael Manoochehri
Longhao Wang
Jordan Shedlock
Daniel Griffin

8.1 Introduction

In Chapter 6 we discussed different types of semantic relationships and contrasted abstract relationships between categories that define a semantic hierarchy like

Meat → is-a → Food

with concrete relationships involving specific people like members of the Simpson family:

Homer Simpson → is-a → Husband

When we make an assertion that a particular instance like Homer Simpson is a member of class, we are *classifying* the instance.

Classification, the systematic assignment of resources to intentional categories, is the focus of this chapter. In Chapter 7, *Categorization: Describing Resource Classes and Types*, we described categories created by people as cognitive and linguistic models for applying prior knowledge and we discussed a set of principles for creating categories and category systems. We explained how cultural categories serve as the foundations upon which individual and institutional categories are based. Institutional categories are most often created in abstract and information-intensive domains where unambiguous and precise categories enable *classification* to be purposeful and principled. Computational categories inherited by supervised learning techniques are usually as interpretable as those created by people, but categories created by unsupervised machine learning techniques are statistical patterns that might or might not be interpretable.

A system of categories and its attendant rules or access methods is typically called a *classification scheme* or just the *classifications*. A system of categories captures the distinctions and relationships among its resources that are most important in a domain and for a particular context of use, creating a reference model or conceptual roadmap for its users. This classification creates the structure and support for the interactions that human or computational agents perform. For example, research libraries and bookstores do not use the same classifications to organize books, but the categories they each use are appropriate for their contrasting types of collections and the different kinds of browsing and searching activities that take place in each context. Likewise, the scientific classifications for animals used by biologists contrast with those used in pet stores because the latter have no need for the precise differentiation enabled by the former.

Navigating This Chapter

Most of the chapter is a survey of topics that span the broad range of how classifications are used in organizing systems. These include enumerative classification (§8.3), faceted classification (§8.4), activity-based classification (§8.5), and computational classification (§8.6). Because classification and standardization are closely related, we also analyze standards and standards making as they apply to organizing systems. Throughout, we observe how personal, institutional, cultural, linguistic, political, religious, and even artistic biases can affect otherwise principled and purposeful classification schemes. We finish the chapter with §8.7 Key Points in Chapter Eight (page 354).

8.1.1 Classification vs. Categorization

Classification requires a system of categories, so not everyone distinguishes classification from categorization. Batley, for example, says classification is "imposing some sort of structure on our understanding of our environment," a vague definition that applies equally well to categorization.

In the discipline of organizing, the definition of classification is narrower and more formal. The contrasts among cultural, individual, and institutional categories in §7.2 The What and Why of Categories (page 269) yield a precise definition of classification: *The systematic assignment of resources to a system of intentional categories, often institutional ones.* This definition highlights the intentionality behind the system of categories, the systematic processes for using them, and implies the greater requirements for *governance* and *maintenance* that are absent for cultural categories and most individual ones.

8.1.2 Classification vs. Tagging

Precise and reliable classification is possible when the shared properties of a collection of resources are used in a principled and systematic manner. This method of classification is essential to satisfy institutional and commercial purposes. However, this degree of rigor might be excessive for personal classifications and for classifications of resources in social or informal contexts.

Instead, a weaker approach to organizing resources is to use any property of a resource and any vocabulary to describe it, regardless of how well it differentiates it from other resources to create a system of categories. This method of organizing resources is most often called *tagging* (§5.2.2.3), but it has also been called *social classification*.

Tagging is often used in personal organizing systems, but is social when it serves goals to convey information, develop a community, or manage reputation. Regardless of its name, however, tagging is popular for organizing and rating photos, websites, email messages, or other web-based resources or web-based descriptions of physical resources like stores and restaurants.

The distinction between classification and tagging was blurred when Thomas Vander Wal coined the term "folksonomy" —combining "folk" and "taxonomy" (which is a classification; see §6.3.1.1 Inclusion (page 231)) —to describe the collection of tags for a particular web site or application. Folksonomies are often displayed in the form of a *tag cloud*, where the frequency with which the tag is used throughout the site determines the size of the text in the tag cloud. The tag cloud emerges through the bottom-up aggregation of user tags and is a statistical construct, rather than a semantic one.

Tagging seems insufficiently principled to be considered classification. Tagging a photo as "red" or "car" is an act of resource description, not classification, be-

cause the other tags that would serve as the alternative classifications are unspecified. Furthermore, when tagging principles are followed at all, they are likely to be idiosyncratic ones that were not pre-determined or arrived at through an analysis of goals and requirements.

Noticeably, some uses of tags treat them as category labels, turning tagging into classification. Many websites and resources encourage users to assign "Like" or "+1" tags to them, and because these tags are pre-defined, they are category choices in an implied classification system; for example, we can consider "Like" as an alternative to a "Not liked enough" category.

When users or communities establish sets of principles to govern their tagging practices, tagging is even more like classification. Such a tagging system can be called a *tagsonomy*, a neologism we have invented to describe more systematic tagging. For example, a tagsonomy could predetermine tags as categories to be assigned to particular contents of a blog post, or specify the level of abstraction and granularity for assigning tags without predetermining them (§7.4 Category Design Issues and Implications (page 298)). Some people use multiple user accounts for the same application to establish distinct personas or contexts (e.g., personal vs. business photo collections) as a way to make their tagsonomies more distinct.

Making these decisions about tagging content and form and applying them in the tagging process transforms an *ad hoc* set of tags into a principled tagsonomy. When tagging is introduced in a business setting, more pragmatic purposes and more systematic tagging—for example, by using tags from lists of departments or products—also tends to create tagsonomic classification.

8.1.3 Classification vs. Physical Arrangement

We have often stressed the principle in the discipline of organizing that logical issues must be separated from implementation issues. (See §1.6 The Concept of "Organizing Principle" (page 41), §5.3.5 Designing the Description Form (page 210), and §6.7 The Implementation Perspective (page 258)) With classification we separate the conceptual act of assigning a resource to a category from the subsequent but often incidental act of putting it in some physical or digital storage location. This focus on the logical essence of classification is elegantly expressed in a definition by Gruenberg: Classification is "a higher order thinking skill requiring the fusion of the naturalist's eye for relationships... with the logician's desire for structured order... the mathematician's compulsion to achieve consistent, predictable results... and the linguist's interest in explicit and tacit expressions of meaning."

Taking a conceptual or cognitive perspective on classification contrasts with much conventional usage in library science, where classification is mostly associated with arranging tangible items on shelves, emphasizing the "parking"

function that realizes the "marking" function of identifying the category to which the resource belongs.

From a library science or collection curation perspective, it seems undeniable that when the resources being classified are physical or tangible things such as books, paintings, animals, or cooking pots, the end result of the classification activity is that some resource has been placed in some physical location. Moreover, the placement of physical resources can be influenced by the physical context in which they are organized. Once placed, the physical context often embodies some aspects of the organization when similar or related resources are arranged in nearby locations. In libraries and bookstores, this adjacency facilitates the serendipitous discovery of resources, as anyone well knows who has found an interesting book by browsing the shelves.

However, once we broaden the scope of organizing to include digital resources, it is clear that we rely on their logical classifications when we interact with them, not whether they reside on a computer in Berkeley or Bangalore. It is better to emphasize that a classification system is foremost a specification for the logical arrangement of resources because there are usually many possible and often arbitrary mappings of logical references to physical locations.

8.1.4 Classification Schemes

A classification scheme is a realization of one or more organizing principles. Physical resources are often classified according to their tangible or perceivable properties. As we discussed in §7.3.2 Single Properties (page 281) and §7.3.3 Multiple Properties (page 283), when properties take on only a small set of discrete values, a classification system naturally emerges in which each category is defined by one property value or some particular combination of property values. Classification schemes in which all possible categories to which resources can be assigned are defined explicitly are *enumerative*. For example, the *enumerative classification* for a personal collection of music recorded on physical media might have categories for CDs, DVDs, vinyl albums, 8-track cartridges, reel-to-reel tape, and tape cassettes; every music resource fits into one and only one of these categories.

When multiple resource properties are considered in a fixed sequence, each property creates another level in the system of categories and the classification scheme is *hierarchical* or *taxonomic*. (See §6.3.1.1 Inclusion (page 231).)

For information resources, their *aboutness* is usually more important than their physical properties. For example, a professor planning a new course might organize candidate articles for the syllabus in a fixed set of categories, one for each potential lecture topic. But it is more challenging to enumerate all the subjects or topics that a larger collection of resources might be about. The Library of Congress Classification (LCC) is a hierarchical and enumerative scheme with

a very detailed set of subject categories because books can be about almost anything. We discuss the LCC more in §8.3 Bibliographic Classification (page 338).

In addition to or instead of their *aboutness,* information resources are sometimes organized using intrinsic properties like author names or creation dates. Our professor might primarily organize his collection of articles by author name, and when he plans a new course, he might put those he selects for the syllabus into a classification system with one category for every scheduled lecture.

Because names and dates can take on a great many values, an organizing principle like *alphabetical* or *chronological* ordering is unlikely to enumerate in advance an explicit category for each possible value. Instead, we can consider these organizing principles as creating an *implicit or latent* classification system in which the categories are generated only as needed. For example, the Q category only exists in an alphabetical scheme if there is a resource whose name starts with Q.

Many resource domains have multiple properties that might be used to define a classification scheme. For example, wine can be classified by type of grape (varietal), color, flavor, price, winemaker, region of origin (appellation), blending style, and other properties. Furthermore, people differ in their knowledge or preferences about these properties; some people choose wine based on its price and varietal, while others studiously compare winemakers and appellations. Each order of considering the properties creates a different hierarchical classification, and using all of them would create a very deep and unwieldy system. Moreover, many different hierarchies might be required to satisfy divergent preferences. An alternative classification scheme for domains like these is *faceted* classification, a type of classification system that takes a set of resource properties and then generates only those categories for combinations that actually occur.

The most common types of facets are enumerative (mutually exclusive); Boolean (yes or no); hierarchical or taxonomic (logical containment); and spectrum (a range of numerical values). We discuss *faceted classification* in detail (in §8.4 Faceted Classification (page 342)) because it is very frequently used in online classifications. Faceted schemes enable easier search and browsing of large resource collections like those for retail sites and museums than hierarchical enumerative schemes.

8.1.5 Classification and Standardization

Classifications impose order on resources. Standards do the same by making distinctions, either implicitly or explicitly, between "standard" and "nonstandard" ways of creating, organizing, and using resources. Classification and standardization are not identical, but they are closely related. Some classifications become standards, and some standards define new classifications. Institutional categories (§7.2.3) are of two broad types.

8.1.5.1 Institutional Taxonomies

Institutional taxonomies are classifications designed to make it more likely that people or computational agents will organize and interact with resources in the same way. Among the thousands of standards published by the *International Organization for Standardization (ISO)* are many institutional taxonomies that govern the classification of resources and products in agriculture, aviation, construction, energy, healthcare, information technology, transportation, and almost every industry sector.

Institutional taxonomies are especially important in libraries and knowledge management. The Dewey Decimal Classification (DDC) and Library of Congress Classification (LCC) enable different libraries to arrange books in the same categories, and the *Diagnostic and Statistical Manual of Mental Disorders (DSM)* in clinical psychology enables different doctors to assign patients to the same diagnostic and insurance categories.

8.1.5.2 Institutional Semantics

Systems of *institutional semantics* offer precisely defined abstractions or *information components* (§4.3.3 Identity and Information Components (page 155)) needed to ensure that information can be efficiently exchanged and used. Organizing systems that use different information models often cannot share and combine information without tedious negotiation and excessive rework.

Standard semantics are especially important in industries or markets that have significant network effects where the value of a product depends on the number of interoperable or compatible products—these include much of the information and service economies.

8.1.5.3 Specifications vs. Standards

Implementing an organizing system of significant scope and complexity in a robust and maintainable fashion requires precise descriptions of the resources it contains, their formats, the classes, relations, structures and collections in which they participate, and the processes that ensure their efficient and effective use. Rigorous descriptions like these are often called "specifications" and there are well-established practices for developing good ones.

There is a subtle but critical distinction between "specifications" and "standards." Any person, firm, or *ad hoc* group of people or firms can create a specification and then use it or attempt to get others to use it. In contrast, a standard is a published specification that is developed and maintained by consensus of all the relevant stakeholders in some domain by following a defined and transparent process, usually under the auspices of a recognized standards organization. In addition, implementations of standards often are subject to conformance tests that establish the completeness and accuracy of the implementation. This means that users can decide either to implement the specification themselves or choose from other conforming implementations.

The additional rigor and transparency when specifications are developed and maintained through a standards process often makes them fairer and gives them more legitimacy. Governments often require or recommend these *de jure* standards, especially those that are "open" or "royalty free" because they are typically supported by multiple vendors, minimizing the cost of adoption and maximizing their longevity.

Despite these important distinctions between "specifications" and "standards," however, in conventional usage "standard" is often simply a synonym for "dominant or widely-adopted specification." These *de facto* standards, in contrast with the *de jure* standards created by standards organizations, are typically created by the dominant firm or firms in an industry, by a new firm that is first to use a new technology or innovative method, or by a non-profit entity like a foundation that focuses on a particular domain.

De facto standards and *ad hoc* standards often co-exist and compete in "standards wars," especially in information-intensive domains and industries with rapid innovation. Standards "wars" tend to occur when different firms or groups of firms develop two or more standards that tend to address the same needs. Not surprisingly, the competing standards are often incompatible on purpose. At first this lets each standard attract customers with features not enabled by the other, but it ends up locking them in by imposing switching costs. Current examples include Google vs. Apple on mobile phones and Kindle versus Apple on ebook readers.

For example, the Dewey Decimal Classification (DDC) is the world's most widely used library classification system, and most people treat it as a standard. In fact, the DDC is proprietary and it is maintained and licensed for use by the Online Computer Library Center (OCLC). Similarly, the DSM is maintained and published by the *American Psychiatric Association (APA)* and it earns the APA many millions of dollars a year.

In contrast, *de jure* standards include the Library of Congress Classification (LCC), developed under the auspices of the US government, the familiar MARC record format used in online library catalogs (ISO 2709), and its American counterpart ANSI Z39.2.

As a result, even though it would be technically correct to argue that "while all standards are specifications, not all specifications are standards," this distinction is hard to maintain in practice.

8.1.5.4 Mandated Classifications

Standards are often imposed by governments to protect the interests of their citizens by coordinating or facilitating activities that might otherwise not be possible or safe. Some of them primarily concern public or product safety and are only tangentially relevant to systems for organizing information. Others are highly relevant, especially those that specify the formats and content of information exchange; many European governments require firms doing business with the government to adopt UBL.

Other government standards that are important in organizing systems are those that express requirements for classification and retention of auditing information for financial activities, such as the *Sarbanes-Oxley Act,* or for non-retention of personal information, such as HIPAA and FERPA.

8.2 Understanding Classification

Classifications arrange resources to support discovery, selection, combination, integration, analysis, and other purposeful activity in every organizing system. A classification of diseases facilitates diagnosis and development of medical procedures, as well as accounting and billing. In addition, classifications facilitate understanding of a domain by highlighting the important resources and relationships in it, supporting the training of people who work in the domain and their acquisition of specialized skills for it.

We consider classification to be systematic when it follows principles that govern the structure of categories and their relationships. However, being systematic and principled does not necessarily ensure that a classification will be unbiased or satisfy all users' requirements. For example, the zoning, environmental, economic development, and political district classifications that overlay

different parts of a city determine the present and future allocation of services and resources, and over time influence whether the city thrives or decays. These classifications reflect tradeoffs and negotiations among numerous participants, including businesses, lobbyists, incumbent politicians, donors to political parties, real estate developers, and others with strong self-interests.

8.2.1 Classification Is Purposeful

Categories often arise naturally, but by definition classifications do not because they are systems of categories that have been intentionally designed for some purpose. Every classification brings together resources that go together, and in doing so differentiates among them. However, bringing resources together would be pointless without reasons for finding, accessing, and interacting with them later.

8.2.1.1 Classifications Are Reference Models

A classification creates a semantic or conceptual roadmap to a domain by highlighting the properties and relationships that distinguish the resources in it. This reference model facilitates learning, comprehension, and use of organizing systems within the domain. Standard classifications like those used in libraries enable people to rely on one system that they can use to locate resources in many libraries. Standard business, job, and product classifications enable the reliable collection, analysis, and interchange of economic data and resources.

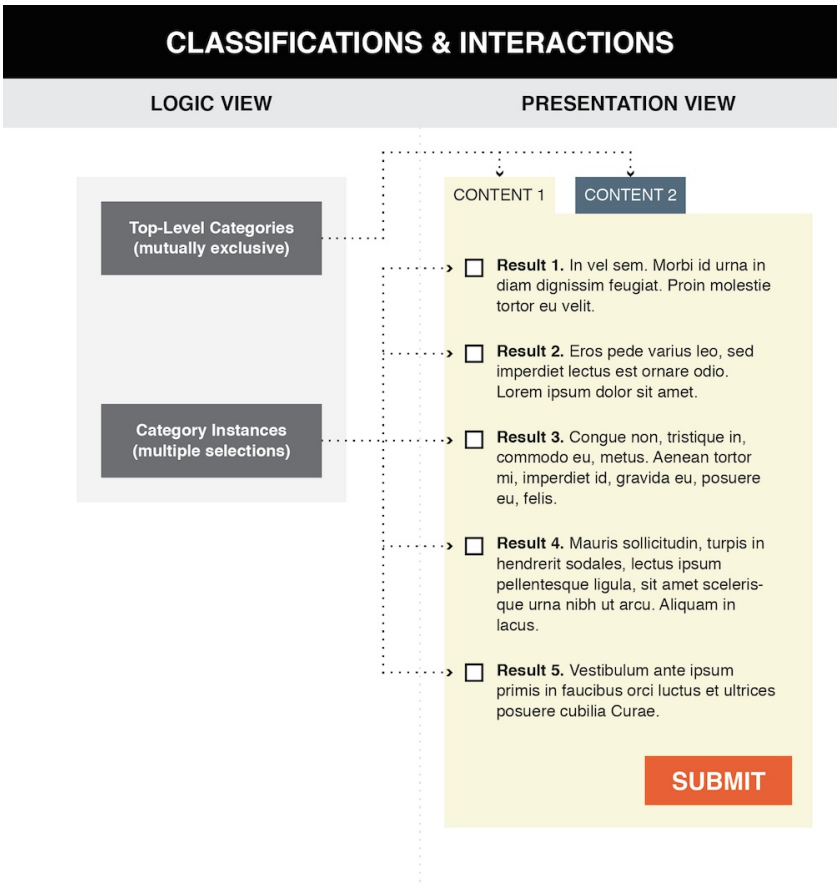
8.2.1.2 Classifications Support Interactions

A classification creates structure in the organizing system that increases the variety and capability of the interactions it can support. With physical resources, classification increases useful co-location; in kitchens, for example, keeping resources that are used together near each other (e.g., baking ingredients) makes cooking and cleanup more efficient (see “activity-based” classification in §8.5).

Classification makes systems more usable when it is manifested in the arrangement of resource descriptions or controls in user interface components like list boxes, tabs, buttons, function menus, and structured lists of search results.

A typical mapping between the logic of a classification scheme and a user interface is illustrated in Figure 8.1, Classification and Interactions.

Figure 8.1. Classification and Interactions.



Good user interface design creates a clear mapping between the logic of a classification scheme and the selection methods and arrangements presented to users. Categories that are mutually exclusive imply different tabs or other visualizations that imply a single selection, for example.

Classification In A Novel User Interface

The meat from animals used as food is classified into numerous "cuts" based on its origin. In the US, these classifications are standardized by the Department of Agriculture to ensure that meat is labeled correctly. The most natural way to convey the classification system is to label the parts of the animal in a diagram, because this binds each logical category to the "user interface."

(Photo by R. Glushko. Taken in 2011 at the Union Square Greenmarket in New York City.)

8.2.2 Classification Is Principled

§7.3 Principles for Creating Categories (page 280) explained principles for creating categories, including enumeration, single properties, multiple properties and hierarchy, probabilistic co-occurrence of properties, theory and goal-based categorization. It logically follows that the principles considered in designing categories are embodied in classifications that use those categories. However, when we say, "classification is principled," we are going further to say that the processes of assigning resources to categories and maintaining the classification scheme over time must also follow principles.

The design and use of a classification system involves many choices about its purposes, scope, scale, intended lifetime, extensibility, and other considerations. Principled classification means that once those design choices are made they should be systematically and consistently followed.

Principled does not necessarily equate to "good," because many of the choices can be arbitrary and others may involve tradeoffs that depend on the nature of the resources, the purposes of the classification, the amount of effort available, the complexity of the domain, and the capabilities of the people doing the classification and of the people using it (see §7.4 Category Design Issues and Implications (page 298)). Every classification system is biased in one way or another (see §8.3 Bibliographic Classification (page 338)).

Consider the classifications of resources in a highly-organized kitchen. (See §12.5 Organizing a Kitchen (page 464)). Tableware, dishes, pots and pans, spices and food provisions, and other resources have dedicated locations determined by a set of intersecting requirements and organizing principles. There is no written specification, and other people organize their kitchens differently.

On the other hand, complex institutional classification systems like those used in libraries or government agencies are implemented with detailed specifications, methods, protocols, and guidelines. The people who apply these methods in the field have studied the protocols in school or they have received extensive on-the-job training to ensure that they apply them correctly, consistently, and in accordance with the specifications and guidelines.

8.2.2.1 Principles Embodied in the Classification Scheme

Some of the most important principles that lead us to say that classification is principled are those that guide the design of the classification scheme in the first place. These principles are fundamental in the discipline of library science but they apply more broadly to other domains.

The *warrant* principle concerns the justification for the choice of categories and the names given to them. The principle of *literary warrant* holds that a classification must be based only on the specific resources that are being classified. In the library context, this *ad hoc* principle that builds a classification from a particular collection principle is often posed in opposition to a more philosophical or epistemological perspective, first articulated by Francis Bacon in the seventeenth century, that a classification should be universal and must handle all knowledge and all possible resources. The principle of *scientific warrant* argues that only the categories recognized by the scientists or experts in a domain should be used in a classification system, and it is often opposed by the principle of *use* or *user warrant*, which chooses categories and descriptive terms according to their frequency of use by everyone, not just experts. With classifications of physical resources like those in a kitchen, we see *object warrant*, where

similar objects are put together, but more frequently the justifying principle will be one of use warrant, where resources are organized based on how they are used.

Starbucks Coffee Sizes: "Anti-User" Warrant?

The Starbucks coffee chain seemingly goes out of its way to confuse its customers by calling the smallest (twelve ounces) of its three coffee sizes the "tall" size, calling its sixteen-ounce size a "*grande*," and calling its largest a "*venti*," which is Italian for twenty (ounces). Outside of Starbucks, something that is "tall" is never also considered "small." Ironically, despite having more than five thousand coffeehouses in over fifty countries, Starbucks has none in Italy where *venti* would be in the local language.

A second principle embodied in a classification scheme concerns the breadth and depth of the category hierarchy. We discussed this in §7.4 Category Design Issues and Implications (page 298) but in the context of classification this principle has additional implications and is framed as the extent to which the scheme is *enumerative* (§8.1.3 Classification vs. Physical Arrangement (page 322)). The decision to classify broadly or precisely depends largely on the variety or heterogeneity of the resources that the system of categories has been designed to organize. Because of the diversity of resources for a sale in a department store, a broad classification is necessary to accommodate everything in the store. Kitchen goods will be grouped together in a few aisles on a single floor. But a specialty kitchen store or a wholesale kitchen supply store for restaurants would classify much more precisely because of the restricted resource domain and the greater expertise of those who want to buy things there. An entire section might be dedicated just to knives, organized by knife type, manufacturer, quality of steel, and other categories that are not used in the kitchen section of the department store.

The precision or enumerativeness of a classification scheme increases the similarity of resources that are assigned to the same category and sharpens the distinctions between resources in different categories. However, when different classifications must be combined, mismatches in their precision or granularity can create challenges (see §10.3 Reorganizing Resources for Interactions (page 406)).

8.2.2.2 Principles for Assigning Resources to Categories

The *uniqueness principle* means the categories in a classification scheme are mutually exclusive. Thus, when a logical concept is assigned to a particular category, it cannot simultaneously be assigned to another category. Resources, however, can be assigned to several categories if they embody several concepts

represented by those different categories. This can present a challenge when a physical storage solution is based on storing resources according to its assigned category in a logical classification system. This is not a serious problem for resource types like technical equipment or tools, for which the properties used to classify them are highly salient, and that have very narrow and predictable contexts of use. It is also not a problem for highly-specialized information resources like scientific research reports or government economic data, which might end up in only one specialized class. However, many resources are inherently more difficult to classify because they have less salient properties or because they have many more possible uses.

We face this kind of problem all the time. For example, should we store a pair of scissors in the kitchen or in the office? One solution is to buy a second pair of scissors so that scissors can be kept in both locations where they are typically used, but this is not practical for many types of resources and this principle would be difficult to apply in a systematic manner.

Many books are about multiple subjects. A self-help book about coping with change in a business setting might reasonably be classified as either about applied psychology or about business. It is not helpful that book titles are often poor clues to their content; *Who Moved My Cheese?* is in fact a self-help book about coping with change in a business setting. Its Library of Congress Classification is BF 637, "Applied Psychology," and at UC Berkeley it is kept in the business school library.

The general solution to satisfying the *uniqueness principle* in library classifications when resources do not clearly fit in a single category is to invent and follow a detailed set of often arbitrary rules. Usually, the primary subject of the book is used for assigning a category, which will then determine the book's place on a shelf.

8.2.2.3 Principles for Maintaining the Classification over Time

Most personal classifications are created in response to a specific situation to solve an emerging organizational challenge. As a consequence, personal classification systems change in an *ad hoc* or opportunistic manner during their limited lifetimes. For example, the classification schemes in your kitchen or closet are deconstructed and disappear when you move and take your possessions to a different house or apartment. Your efforts to re-implement the classifications will be influenced by the configuration of shelves and cabinets in your new residence, so they will not be exactly the same.

In contrast, the institutional classification schemes for many library resources, culturally or scientifically-important artifacts, and much of the information created or collected by businesses, governments and researchers might have useful lives of decades or centuries. Classification systems like these can only be

changed incrementally to avoid disruption of the work flows of the organization. We described maintaining resources as an activity in all organizing systems (§3.5 Maintaining Resources (page 116)) and the issues of persistence, effectivity, authenticity, and provenance that emerge with resources over time (§4.5 Resources over Time (page 167)). Much of this previous discussion applies in a straightforward manner to maintaining classifications over time.

However, some additional issues arise with classifications over time. The warrant principle (§8.2.2.1) implicitly treats the justification for designing and naming categories as a one-time decision. This is reasonable if you are organizing a collection of bibliographic resources or common types of physical resources like printed books, clothing or butterflies. However, in domains where the resources are active, change their state or implementation, or otherwise have a probabilistic character it might be necessary to revisit warrant and the decisions based on it from time to time. Put another way, if the world that you are sampling from or describing has some randomness or change in it, the categories and descriptions you imposed on it probably need to change as well. It often happens that the meaning of an underlying category can change, along with its relative and absolute importance with respect to the other categories in the classification system. Categories sometimes change slowly, but they can also change quickly and radically as a result of technological, process, or geopolitical innovation or events. Entirely new types of resources and bodies of knowledge can appear in a short time. Consider what the categories of "travel," "entertainment," "computing," and "communication" mean today compared to just a decade or two ago.

Changes in the meaning of the categories in a classification threaten its *integrity*, the principle that categories should not move within the structure of the classification system. One way to maintain integrity while adapting to the dynamic and changing nature of knowledge is to define a new version of a classification system while allowing earlier ones to persist, which preserves resource assignments in the previous version of the classification system while allowing it to change in the new one. If we adopt a logical perspective on classification (§8.1.2 Classification vs. Tagging (page 321)) that dissociates the conceptual assignment of resources to categories from their physical arrangement, there is no reason why a resource cannot have contrasting category assignments in different versions of a classification.

However, the conventional library with collections of physical resources cannot easily abandon its requirement to use a classification to arrange books on shelves in specific places so they can be located, checked out, and returned to the same location.

A related principle about maintaining classifications over time is *flexibility*, the degree to which the classification can accommodate new categories. Computer

scientists typically describe this principle as *extensibility*, and library scientists sometimes describe it as *hospitality*. In any case the concern is the same and we are all familiar with it. When you buy a bookshelf, clothes wardrobe, file cabinet, or computer, it makes sense to buy one that has some extra space to accommodate the books, clothes, or files you will acquire over some future time frame. As with other choices that need to be made about organizing systems, how much extra space and "organizing room" you will acquire involves numerous tradeoffs.

Classification schemes can increase their flexibility by creating extra "logical space" when they are defined. Library classifications accomplish this by using naming or numbering schemes for classification that can be extended easily to create new subcategories. Classification schemes in information systems can also anticipate the evolution of document or database schemas.

8.2.3 Classification Is Biased

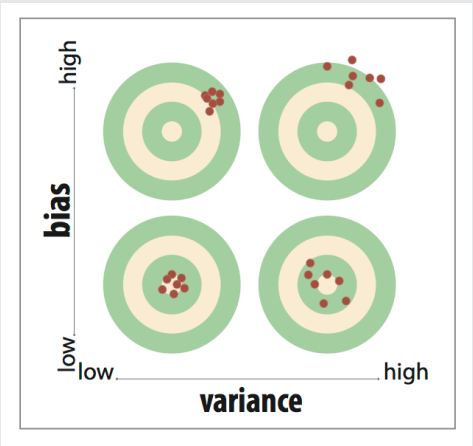
The discipline of organizing is fundamentally about choices of properties and principles for describing and arranging resources. We discussed choices about describing resources in §5.3 The Process of Describing Resources (page 188), choices for creating resource categories in §7.3 Principles for Creating Categories (page 280), and choices for creating classifications in this chapter. The choices made reflect the purposes, experiences, professions, politics, values, and other characteristics and preferences of the people making them. As a result, every system of classification is biased because it takes a point of view that is a composite of all of these influences.

But first we need to point out that there are at least two quite different senses of "bias" that people reading this book are likely to encounter. The colloquial sense of bias we discuss in this section reflects value-based decisions in organizing systems that implicitly or explicitly favor some interactions or users over others. In contrast, statistical bias is systematic error or distortion in a measurement. (See the sidebar, Statistical Bias and Variance (page 335).)

The claim that classification is biased might seem surprising, because many classification systems are formal and institutional, created by governments or firms participating in standards organizations. We expect these classifications to be impartial and objective. However, consider the classification of people as "employed" or "unemployed." Many people think that any employable person who is not currently employed would be counted as unemployed. But the US government's Department of Labor only counts someone as unemployed if they have actively looked for work in the past month, effectively removing anyone who has given up on finding work from the unemployed category by assigning them to a "discouraged worker" category. In 2012 this classification scheme allowed the government to report that unemployment was about 8% and falling,

Bias and Variance on Dartboards

Precise and accurate dart throws demonstrate low bias and low variance (lower left in the figure). Precise but inaccurate darts reflect high bias and low variance (upper left). Imprecise but accurate ones have low bias but high variance (lower right). Finally, a lack of accuracy and precision shows both high bias and high variance (upper right).



Statistical Bias and Variance

Statistical bias is the systematic error in measurements introduced by miscalibration of the measurement instrument, by ineffective measurement techniques, an algorithm that makes incorrect assumptions, or some environment interference, all of which distort the measured value in a predictable way. Measurement bias contrasts with the variability or variance of a measurement, the amount of dispersion around an average or expected value, most often due to random factors. Some variance arises because the property being measured is not the same for all instances, as we would expect for measurements of the weight of a random sample of people, or in the set of tags or topics assigned to a random sample of news articles by people or algorithms. By analyzing a large enough set of instances it is possible to determine the most likely values of the property and also to estimate the amount of random error.

High variance in the measurements for a sample of resources when we expect all of them to have more similar values can be a quality problem. High bias, on the other hand, might be less of a quality problem, because systematic sources of inaccuracy might be easier to correct.

when in fact it was closer to 20% and rising. The political implications of this classification are substantial.

Classification bias is often intentionally or unintentionally shown in data visualizations, including choropleth maps, in which map regions are colored, patterned, or otherwise distinguished according to a statistical variable being displayed on the map. Choropleths are commonly used to display election results, with the districts or states won by each candidate shown in different colors; in the United States, the

California Election Map: A "Blue" State with "Red" Counties

convention is to show those won by Democratic Party candidates in blue, and those won by Republicans in Red. These election choropleths are often misleading because coloring an entire state in the winner's colors ignores population density and the regional concentrations of votes that differ from the majority. California voters are reliably "blue" as a whole, but as you can see in the nearby figure with election results divided by county, this majority is amassed in the large cities along the coast, and inland and rural counties are more reliably "red" in their voting.

A more subtle way in which choropleths encode bias reflects the decisions made to organize the data into the categories that are represented by different colors or patterns. Choropleth categories might present data divided into equal range intervals, into sets with the same number of observations, or into categories that reflect clusters or natural breaks in the observed data. Small changes in the data ranges or proportions that are then assigned to each category can communicate entirely different stories with the same data. To learn "how to lie with maps" or how to prevent being lied to, refer to the classic book with that title by Mark Monmonier.

Bowker and Star have written extensively about biases in classification systems but acknowledge that many people do not see them:

> *Information scientists work every day on the design, delegation and choice of classification systems and standards, yet few see them as artifacts embodying moral and aesthetic choices that in turn craft people's identities, aspirations and dignity.*
>
> — (Bowker and Star 2000)

Bowker and Star describe many examples where seemingly neutral and benign classifications implement controversial assumptions. A striking example is found in the ethnic classifications of the United States Census and the categories to which US residents are required to assign themselves. These categories have changed nearly every decade since the first census in 1790 and strongly reflect

political goals, prevailing cultural sensitivities or lack thereof, and non-scientific considerations. Some recent changes included a "multi-racial" category, which some people viewed as empowering, but which was attacked by African-American and Hispanic civil rights groups as diluting their power.

A more positive way to think about bias in classification is that the choices made in an organizing system about resource selection, description, and arrangement come together to convey the values of the organizers. This makes a classification a rhetorical or communicative vehicle for establishing credibility and trust with those who interact with the resources in the classification. Seen in this light, an objective or neutral classification is not only unrealistic as a goal; it may also consume valuable time and energy when instead it might be more desirable to seize the opportunity to interpret the resources in a creative way to communicate a particular message to a particular user group. Melanie Feinberg makes the point that "fair trade" or "green" supermarkets differentiate themselves by a relatively small proportion of the goods they offer compared with ordinary stores, but these particular items signal the values that their customers care most about.

Bias is clearly evident in the most widely used bibliographic classifications, the Library of Congress and the Dewey Decimal, which we discuss next.

8.3 Bibliographic Classification

Much of our thinking about classification comes from the bibliographic domain. Libraries and the classification systems for the resources they contain have been evolving for millennia, shaped by the intellectual, social, and technological conditions of the societies that created them. As early as the third millennium BCE, there were enough written documents—papyrus scrolls or clay tablets—that the need arose to organize them. Some of the first attempts, by Mesopotamian scribes, were simple lists of documents in no particular order. The ancient Greeks, Romans, and Chinese created more principled systems, both sorting works by features such as language and alphabetical order, and placing them into semantically significant categories such as topic or genre. Medieval European libraries were tightly focused on Christian theology, but as secular books and readers proliferated thanks to new technologies and increased literacy, bibliographic classifications grew broader and more complex to accommodate them. Modern classification systems are highly nuanced systems designed to encompass all knowledge; however, they retain some of the same features and biases of their forebears.

We will briefly describe the most important systems for bibliographic classification, especially the Dewey Decimal Classification (DDC) and *Library of Congress Classification (LCC)* systems. However, there are several important ways in which bibliographic classification is distinctive and we will discuss those first:

Scale, Complexity, and Degree of Standardization:
Department stores and supermarkets typically offer tens of thousands of different items (as measured by the number of "stock keeping units" or SKUs), and popular online commerce sites like Amazon.com and eBay are of similar scale. However, the standard product classification system for supermarkets has only about 300 categories. The classifications for online stores are typically deeper than those for physical stores, but they are highly idiosyncratic and non-standard. In contrast, scores of university libraries have five million or more distinct items in their collections, and they almost all use the same standard bibliographic classification system that has about 300,000 distinct categories.

Legacy of Physical Arrangement, User Access, and Re-Shelving:
A corollary to the previous one that distinguishes bibliographic classification systems is that they have long been shaped and continue to be shaped by the legacy of physical arrangement, user access to the storage locations, and re-shelving that they support. These requirements constrain the evolution and extensibility of bibliographic classifications, making them less able to keep pace with changing concepts and new bodies of knowledge. Amazon classifies the products it sells in huge warehouses, but its customers do not have to pick out their purchases there, and most goods never return to the warehouse. Amazon can add new product categories and manage the resources in warehouses far more easily than libraries can.

With digital libraries, constraints of scale and physical arrangement are substantially eliminated, because the storage location is hidden from the user and the resources do not need to be returned and re-shelved. However, when users can search the entire content of the library, as they have learned to expect from the web, they are less likely to use the bibliographic classification systems that have painstakingly been applied to the library's resources.

8.3.1 The Dewey Decimal Classification

The Dewey Decimal Classification (DDC) is the world's most widely used bibliographic system, applied to books in over 200,000 libraries in 135 countries. It is a proprietary and *de facto* standard, and it must be licensed for use from the Online Computer Library Center (OCLC).

In 1876, Melvil Dewey invented the DDC when he was hired to manage the Amherst College library immediately after graduating. Dewey was inspired by Bacon's attempt to create a universal classification for all knowledge and considered the DDC as a numerical overlay on Bacon with 10 main classes, each divided into 10 more, and so on. Despite his explicit rejection of literary warrant, however, Dewey's classification was strongly influenced by the existing Amherst

Figure 8.2. "Religion" in Dewey Decimal Classification.

```
200 Religion
  210 Natural Theology
  220 Bible
  230 Christian theology
  240 Christian moral and devotional theology
  250 Christian orders and local church
  260 Christian social theology
  270 Christian church history
  280 Christian sects and denominations
  290 Other religions
```

collection, which reflected Amherst's focus on the time on the "education of indigent young men of piety and talents for the Christian ministry."

The resulting nineteenth-century Western bias in the DDC's classification of religion seems almost startling today, where it persists in the 23rd revision (see Figure 8.2, "Religion" in Dewey Decimal Classification.). "Religion" is one of the 10 main classes, the 200 class, with nine subclasses, Six of these nine subclasses are topics with "Christian" in the name; one class is for the *Bible* alone; and another section is entitled "Natural theology." Everything else related to the world's many religions is lumped under 290, "Other religions."

The notational simplicity of a decimal system makes the DDC easy to use and easy to subdivide existing categories, So-called subdivision tables allow facets for language, geography or format to be added to many classes, making the classification more specific. But the overall system is not very hospitable to new areas of knowledge.

8.3.2 The Library of Congress Classification

The US Library of Congress is the largest library in the world today, but it got off to a bad start after being established in 1800. In 1814, during the War of 1812, British troops burned down the US Capitol building where the library was located and the 3000 books in the collection went up in flames. The library was restarted a year later when Congress purchased the personal library of former president Thomas Jefferson, which was over twice the size of the collection that the British burned. Jefferson was a deeply intellectual person, and unlike the narrow historical and legal collection of the original library, Jefferson's library reflected his "comprehensive interests in philosophy, history, geography, science, and literature, as well as political and legal treatises."

Restarting the Library of Congress around Jefferson's personal collection and classification had an interesting implication. When Herbert Putnam formally created the Library of Congress Classification (LCC) in 1897, he meant it not as

Figure 8.3. Top Level Categories in the Library of Congress Classification.

A – GENERAL WORKS
B – PHILOSOPHY. PSYCHOLOGY. RELIGION
C – AUXILLARY SCIENCES OF HISTORY (GENERAL)
D – WORLD HISTORY (EXCEPT AMERICAN HISTORY)
E – HISTORY: AMERICA
F – HISTORY: AMERICA
G – GEOGRAPHY. ANTHROPOLOGY. RECREATION
H – SOCIAL SCIENCE
J – POLITICAL SCIENCE
K – LAW
L – EDUCATION
M – MUSIC
N – FINE ARTS
P – LANGUAGE AND LITERATURE
Q – SCIENCE
R – MEDICINE
S – AGRICULTURE
T – TECHNOLOGY
U – MILITARY SCIENCE
V – NAVAL SCIENCE
Z – BIBLIOGRAPHY. LIBRARY SCIENCE

a way to organize all the world's knowledge, but to provide a practical way to organize and later locate items within the Library of Congress's collection. However, despite Putnam's commitment to literary warrant, the breadth of Jefferson's collection made the LCC more intellectually ambitious than it might otherwise had been, and probably contributed to its dominant adoption in university libraries.

The LCC has 21 top-level categories, identified by letters instead of using numbers like the DDC (see Figure 8.3, Top Level Categories in the Library of Congress Classification.). Each top-level category is divided into about 10-20 subclasses, each of which is further subdivided. The complete LCC and supporting information takes up 41 printed volumes.

Bias is apparent in the LCC as it is in the DDC, but is somewhat more subtle. A library for the US emphasizes its own history. "Naval science" was vastly more important in the 1800s when it was given its own top level category, separated from other resources about "Military science" (which had a subclass for "Cavalry").

The LCC is highly enumerative, and along with the uniqueness principle, this creates distortions over time and sometimes requires contortions to incorporate

new disciplines. For example, it might seem odd today that a discipline as broad and important as computer science does not have its own second level category under the Q category of science, but because computer science was first taught in math departments, the LCC has it as the QA76 subclass of mathematics, which is QA.

8.3.3 The BISAC Classification

A very different approach to bibliographic classification is represented in the *Book Industry Standards Advisory Committee classification (BISAC)*. BISAC is developed by the *Book Industry Study Group (BISG)*, a non-profit industry association that "develops, maintains, and promotes standards and best practices that enable the book industry to conduct business more efficiently." The BISAC classification system is used by many of the major businesses within the North American book industry, including Amazon, Baker & Taylor, Barnes & Noble, Bookscan, Booksense, Bowker, Indigo, Ingram and most major publishers.

The BISAC classifications are used by publishers to suggest to booksellers how a book should be classified in physical and online bookstores. Because of its commercial and consumer focus, BISAC follows a principle of use warrant, and its categories are biased toward common language usage and popular culture. Some top-level BISAC categories, including Law, Medicine, Music, and Philosophy, are also top-level categories in the LCC. However, BISAC also has top-level categories for Comics & Graphic Novels. Cooking, Pets, and True Crime.

The differences between BISAC and the LCC are understandable because they are used for completely different purposes and generally have little need to come into contact. This changed in 2004, when Google began its ambitious project to digitize the majority of the world's books. (See the sidebar, What Is a Library? (page 37)). To the dismay of many people in the library and academic community, Google initially classified books using BISAC rather than the LCC.

In addition, some new public libraries have adopted BISAC rather than the DDC because they feel the former makes the library friendlier to its users. Some librarians believe that their online catalogs need to be more like web search engines, so a less precise classification that uses more familiar category terms seems like a good choice.

8.4 Faceted Classification

We have noted several times that strictly enumerative classifications constrain how resources are assigned to categories and how the classification can evolve over time. *Faceted classifications* are an alternative that overcome some of these limitations. In a *faceted classification* system, each resource is described using properties from multiple facets, but a person searching for resources does

not need to consider all of the properties (and consequently the facets) and does not need to consider them in a fixed order, which an enumerative hierarchical classification requires.

Faceted classifications are especially useful in web user interfaces for online shopping or for browsing a large and heterogeneous museum collection. The process of considering facets in any order and ignoring those that are not relevant implies a dynamic organizational structure that makes selection both flexible and efficient. We can best illustrate these advantages with a shopping example in a domain that we are familiar with from §7.3.3.

If a department store offers shirts in various styles, colors, sizes, brands, and prices, shoppers might want to search and sort through them using properties from these facets in any order. However, in a physical store, this is not possible because the shirts must be arranged in actual locations in the store, with dress shirts in one area, work shirts in another, and so on.

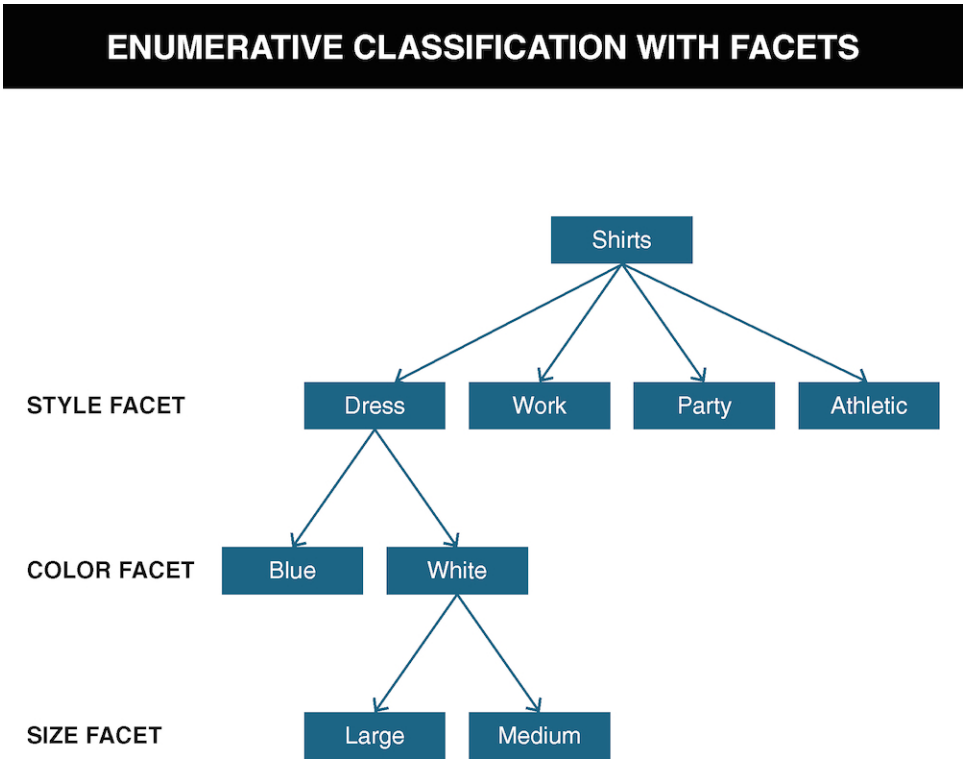
Assume that the shirt store has shirts in four styles: dress shirts, work shirts, party shirts, and athletic shirts. The dress shirts come in white and blue, the work shirts in white and brown, and the party and athletic shirts come in white, blue, brown, and red. White dress shirts come in large and medium sizes.

Suppose we are looking for a white dress shirt in a large size. We can think of this desired shirt in two equivalent ways, either as a member of a category of "large white dress shirts" or a shirt with "dress," "white," and "large" values on style, color, and size facets. Because of the way the shirts are arranged in the physical store, our search process has to follow a hierarchical structure of categories. We go to the dress shirt section, find white shirts, and then look for a large one. This process corresponds to the hierarchy shown in Figure 8.4, Enumerative Classification with Style Facet Followed by Color Facet.

Although unlikely, a store might choose to organize its shirts by color. In our search for a "white dress shirt in a large size," if we consider the color first, because shirts come in four colors, there are four color categories to choose from. When we choose the white shirts, there is no category for work shirts because there are no work shirts that come in white. We then choose the dress shirts, and then finally find the large one. (Figure 8.5, Enumerative Classification with Color Facet Followed by Style Facet.)

This department store example shows that for a physical organization, one property facet guides the localization of resources; all other facets are subordinated under the primary organizing property. In hierarchical enumerative classifications, this means that the primary organizing facet determines the primary form of access. The shirts are either organized by style and then color, or by color then style, which enforces an inflexible query strategy (style first or color first).

Figure 8.4. Enumerative Classification with Style Facet Followed by Color Facet.

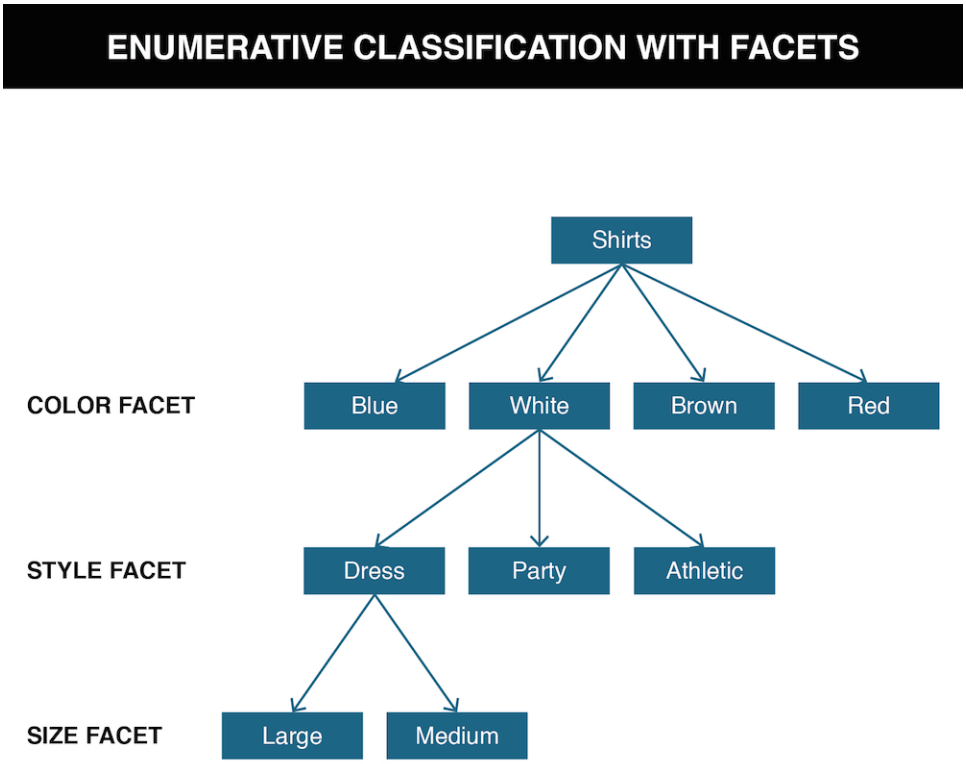


In an enumerative classification system the order of the facets determines the classification hierarchy. For example, a store might classify shirts first using a style facet, next with a color facet, and finally with a size facet. This ordering could result in two piles of dress shirts, one blue and one white, in which each pile contains shirts of large and medium sizes.

We can enumerate all the properties needed to assign resources appropriately, but we create the categories (i.e., union of properties from different facets) only as needed to sort resources with a particular combination of properties.

An additional aspect of the flexibility of faceted classification is that a facet can be left out of a resource description if it is not needed or appropriate. For example, because party shirts are often multi-colored with exotic patterns, it is not that useful to describe their color. Likewise, certain types of athletic shirts might be very loose-fitting, and as a result not be given a size description, but their color is important because it is tied to a particular team. Figure 8.6,

Figure 8.5. Enumerative Classification with Color Facet Followed by Style Facet.

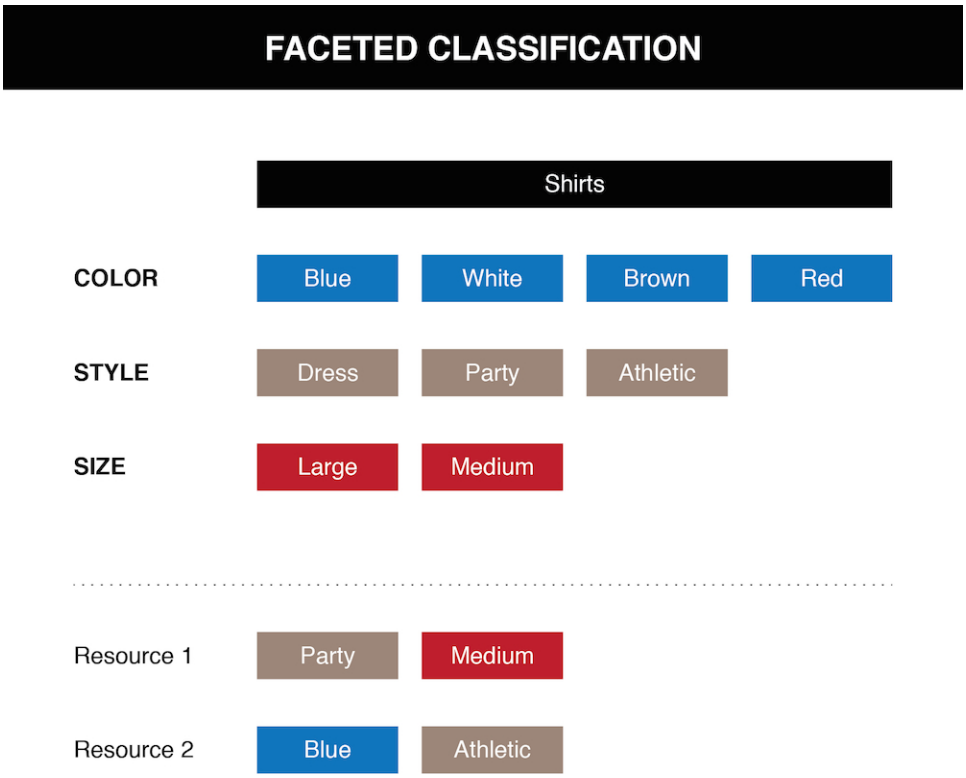


An alternative ordering of the same shirt facets changes the classification hierarchy. If the first facet considered is color, style is next, and finally size, this ordering could result in two piles of white shirts, one for dress shirts and one for athletic shirts, in which each pile contains shirts of large and medium sizes.

Faceted Classification. shows how these two resource types can be classified with the faceted Shirt classification. Resource 1 describes a party shirt in medium; resource 2 describes an athletic shirt in blue without information about size.

A faceted classification scheme like that shown in Figure 8.6, Faceted Classification. eliminates the requirement for predetermining a combination and ordering of facets like those in Figure 8.4, Enumerative Classification with Style Facet Followed by Color Facet. and Figure 8.5, Enumerative Classification with Color Facet Followed by Style Facet. Instead, imagine a shirt store where you decide when

Figure 8.6. Faceted Classification.



In a pure faceted classification, not every facet needs to apply to every resource, and there is no requirement for a predetermined order in which the facets are considered.

you begin shopping which facets are important to you ("show me all the medium party shirts," "show me the blue athletic shirts") instead of having to adhere to whatever predetermined (pre-combined) enumerative classification the store invented. In a digital organizing system, faceted classification enables highly flexible access because prioritizing different facets can dynamically reorganize how the collection is presented.

8.4.1 Foundations for Faceted Classification

In library and information science texts it is common to credit the idea of faceted classification to S.R. Ranganathan, a Hindu mathematician working as a librarian. Ranganathan had an almost mystical motivation to classify everything in the universe with a single classification system and notation, considering it his dharma (the closest translation in English would be "fundamental duty" or "destiny"). Facing the limitations of Dewey's system, where an item's essence had to first be identified and then the item assigned to a category based on that essence, Ranganathan believed that all bibliographic resources could be organized around a more abstract variety of aspects.

In 1933 Ranganathan proposed that a set of five facets applied to all knowledge:

***P**ersonality*
: The type of thing.

***M**atter*
: The constituent material of the thing.

***E**nergy*
: The action or activity of the thing.

***S**pace*
: Where the thing occurs.

***T**ime*
: When the thing occurs.

This classification system is known as colon classification (or PMEST) because the notation used for resource identifiers uses a colon to separate the values on each facet. These values come from tables of categories and subcategories, making the call number very compact. Colon classification is most commonly used in libraries in India.

Ranganathan deserves credit for implementing the first faceted classification system, but people other than librarians generally credit the idea to Nicolas de Condorcet, a French mathematician and philosopher. About 140 years before Ranganathan, Condorcet was concerned that "systems of classification that imposed a given interpretation upon Nature... represented an insufferable obstacle to... scientific advance." Condorcet thus proposed a flexible classification scheme for "arranging a large number of subjects in a system so that we may straightway grasp their relations, quickly perceive their combinations, and readily form new combinations."

Faceted classification is most commonly used in narrow domains, each with its own specific facets. This makes intuitive sense because even if resources can be distinguished with a general classification, doing so requires lengthy notations, and it is much harder to add to a general classification than to a classification created specifically for a single subject area. We could probably describe shirts using the PMEST facets, but style, color, and size seem more natural.

8.4.2 Faceted Classification in Description

Elaine Svenonius defines facets as "groupings of terms obtained by the first division of a subject discipline into homogeneous or semantically cohesive categories." The relationships between these facets results in a controlled vocabulary (§4.1.2) governing the resources we are organizing. From this controlled vocabulary we can generate many descriptions that are complex but formally structured, enabling us to describe things for which terms do not yet exist.

Getty's Art & Architecture Thesaurus (AAT) is a robust and widely used *controlled vocabulary* consisting of generic terms to describe artifacts, objects, places and concepts in the domains of "art, architecture, and material culture."

AAT is a thesaurus with a faceted hierarchical structure. The AAT's facets are "conceptually organized in a scheme that proceeds from abstract concepts to concrete, physical artifacts:"

Associated Concepts
: Concepts, philosophical and critical theory, and phenomena, such as "love" and "nihilism."

Physical Attributes
: Material characteristics that can be measured and perceived, like "height" and "flexibility."

Styles and Periods
: Artistic and architectural eras and stylistic groupings, such as "Renaissance" and "Dada."

Agents
: Basically, people and the various groups and organizations with which they identify, whether based on physical, mental, socio-economic, or political characteristics—e.g., "stonemasons" or "socialists."

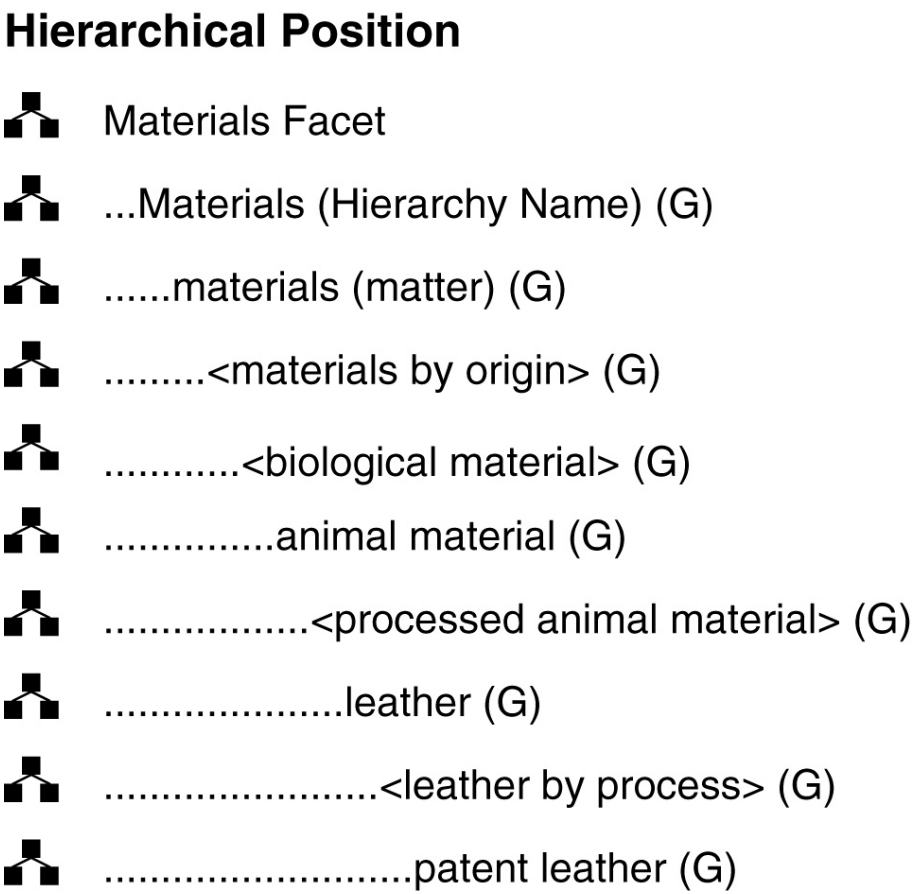
Activities
: Actions, processes, and occurrences, such as "body painting" and "drawing." These are different from the "Objects" facet, which may also contain "body painting," in terms of the actual work itself, not the creation process.

Materials
: Concerned with the actual substance of which a work is made, like “metal” or “bleach.” “Materials” differ from “Physical Attributes” in that the latter is more abstract than the former.

Objects
: The largest facet, *objects* contains the actual works, like “sandcastles” and “screen prints.”

Within each facet is a strict hierarchical structure drilling down from broad term to very specific instance.

Figure 8.7. “Patent Leather” in the Art & Architecture Thesaurus.



The Art and Architecture Thesaurus has a faceted hierarchical structure.

Figure 8.7, “Patent Leather” in the Art & Architecture Thesaurus. shows how a particular instance may be described on a number of dimensions for the purpose of organizing the item and retrieving information about it. And by using a standard *controlled vocabulary*, catalogers and indexers make it easier for users to understand and adapt to the way things are organized for the purpose of finding them.

8.4.3 A Classification for Facets

There are four major types of facets.

Enumerative facets
: Have mutually exclusive possible values. In our online shirt store, “Style” is an enumerative facet whose values are “dress,” “work,” “party,” and “athletic.”

Boolean facets
: Take on one of two values, yes (true) or no (false) along some dimension or property. On a sportswear website, “Waterproof” would be a Boolean facet because an item of clothing is either waterproof or it is not.

Hierarchical facets
: Organize resources by logical inclusion (§6.3.1.1). At Williams-Sonoma’s website, the top-level facet includes “Cookware,” “Cooks’ Tools,” and “Cutlery.” At wine.com the “Region” facet has values for “US,” “Old World,” and “New World,” each of which is further divided geographically. Also see *taxonomic facets*.

Spectrum facets
: Assume a range of numerical values with a defined minimum and maximum. Price and date are common spectrum facets. The ranges are often modeled as mutually exclusive regions (potential price facet values might include “$0—$49,” “$50—$99,” and “$100—$149”).

8.4.4 Designing a Faceted Classification System

It is important to be systematic and principled when designing a faceted classification. In some respects the process and design concerns overlap with those for describing resources, and much of the advice in §5.3 The Process of Describing Resources (page 188) is relevant here.

8.4.4.1 Design Process for Faceted Classification

We advocate a five step process for designing a faceted classification system.

1. Define the purposes of the classification (§5.3.2 Determining the Purposes (page 194), §8.2.1 Classification Is Purposeful (page 328)) and specify the collection of concepts or resources to be classified.
2. For each facet, determine its logical type (§8.4.3 A Classification for Facets (page 350)) and possible values. Specify the order of the values for each facet so that they make sense to users; useful orderings are alphabetical, chronological, procedural, size, most popular to least popular, simple to complex, and geographical or topological.
3. Analyze and describe a representative sample of resource instances to identify properties or dimensions as candidate facets (See §5.3.3 Identifying Properties (page 201)).
4. Examine the relationships between the facets to create sub-facets if necessary. Determine how the facets will be combined to generate the classifications.
5. Test the classification on new instances, and revise the facets, facet values, and facet grammar as needed.

8.4.4.2 Design Principles and Pragmatics

Here is some more specific advice about selecting and designing facets and facet values:

Orthogonality
: Facets should be independent dimensions, so a resource can have values of all of them while only having one value on each of them. In an online kitchen store, one facet might be “Product” and another might be “Brand.” A particular item might be classified as a “Saucepan” in the “Product” facet and as “Calphalon” in the “Brand” one. Other saucepans might have other brands, and other Calphalon products might not be saucepans, because Product and Brand are orthogonal.

Semantic Balance
: Top-level facets should be the properties that best differentiate the resources in the classification domain. The values should be of equal semantic scope so that resources are distributed among the subcategories. Subfacets of “Cookware” like “Sauciers and Saucepans” and “Roasters and Brasiers” are semantically balanced as they are both named and grouped by cooking activity.

Coverage
: The values of a facet should be able of classifying all instances within the intended scope.

Scalability
: Facet values must accommodate potential additions to the set of instances. Including an "Other" value is an easy way to ensure that a facet is flexible and hospitable to new instances, but it not desirable if all new instances will be assigned that value.

Objectivity
: Although every classification has an explicit or implicit bias (§8.2.3 Classification Is Biased (page 335)), facets and facet values should be as unambiguous and concrete as possible to enable reliable classification of instances.

Normativity
: To make a faceted classification as useful by as many people as possible, the terms used for facets and facet values should not be idiosyncratic, metaphorical, or require special knowledge to interpret.

As we will see in §8.6 Computational Classification (page 353), classification can sometimes be done by computers rather than by people. Computer algorithms can analyze resource properties and descriptions to identify dimensions on which resources differ and the most frequent descriptive terms, which can then be used to design a faceted classification scheme. Resources can then be assigned to the appropriate categories, either without human intervention or in collaboration with a human who trains the algorithm with classified instances.

8.5 Classification by Activity Structure

Institutional classification systems are often strongly hierarchical and taxonomic because their many users come to them for diverse purposes, making a context-free or semantic organization the most appropriate. However, in narrow domains that offer a more limited variety of uses it can be much more effective to classify resources according to the tasks or activities they support. A task or activity-based classification system is called a *taskonomy*, a term invented by anthropologists Janet Dougherty and Charles Keller after their ethnographic study of how blacksmiths organized their tools. Instead of keeping things together according to their semantic relationships in what Donald Norman called "hardware store organization," the blacksmiths arranged tools in locations where they were used— "fire tools," "stump tools," "drill press rack tools," and so on.

Personal organizing systems are often taskonomic. Think about the way you cook when you are following a recipe. Do you first retrieve all the ingredients from their storage places, and arrange them in activity-based groups in the preparation area?

Looking at the relationship between tasks and tools in this way can help a cook determine the best way to organize tools in a kitchen. Cutting items would nec-

essarily be kept together near a prep area; having to run across the kitchen to another area where a poultry knife is kept with, say, chicken broth would be detrimental to the cook's workflow. It would make far more sense to have all of the items for the task of cutting in a single area.

The intentional arrangement of tools in a working kitchen might look something like Table 8.1:

Table 8.1. A cook's taskonomy

Prep	Oven	Stove
Poultry knife	Oven mitts	Pots and pans
Paring knife	Baking sheets	Wooden spoons
Vegetable knife	Aluminum foil	Wok
Cutting board	Parchment paper	
	Roasting pan	

Stop and Think: Office Taskonomy

Think about your personal office space. It may be an interesting hybrid space—it probably contains documents that could be classified in a hierarchical system, but it is also a work space that could lend itself to "taskonomy" organization. Which does it more closely resemble? How have any conflicts between hierarchy and "taskonomy" been resolved?

8.6 Computational Classification

Because of its importance, ubiquity, and ease of processing by computers, it should not be surprising that a great many computational classification problems involve text. Some of these problems are relatively simple, like identifying the language in which a text is written, which is solved by comparing the probability of one, two, and three character-long contiguous strings in the text against their probabilities in different languages. For example, in English the most likely strings are "the", "and", "to", "of", "a", "in", and so on. But if the most likely strings are "der", "die", "und", and "den" the text is German and if they are "de", "la", "que", "el", and "en" the text is Spanish.

More challenging text classification problems arise when more features are required to describe each instance being classified and where the features are less predictable. The unknown author of a document can sometimes be identified by analyzing other documents known to be written by him to identify a set of features like word frequency, phrase structure, and sentence length that create a "writeprint" analogous to a fingerprint that uniquely identifies him. This kind of analysis was used in 2013 to determine that *Harry Potter* author J. K.

Rowling had written a crime fiction novel entitled *The Cuckoo's Calling* under the pseudonym Robert Galbraith.

Another challenging text classification problem is sentiment analysis, determining whether a text has a positive or negative opinion about some topic. Much academic and commercial research has been conducted to understand the sentiment of Twitter tweets, Facebook posts, email sent to customer support applications, and other similar contexts. Sentiment analysis is hard because messages are often short so there is not much to analyze, and because and because sarcasm, slang, clichés, and cultural norms obscure the content needed to make the classification.

A crucial consideration whenever supervised learning is used to train a classifier is ensuring that the training set is appropriate. If we were training a classifier to detect spam messages using email from the year 2000, the topics of the emails, the words they contain, and perhaps even the language they are written in would be substantially different than messages from this year. Up to date training data is especially important for the classification algorithms used by Twitter, Facebook, YouTube, and similar social sites that classify and recommend content based on popularity trends.

How a computational classifier "learns" depends on the specific machine learning algorithm. Decision trees, Naive Bayes, support vector machines, and neural net approaches were briefly described in §7.5 Implementing Categories (page 302).

Stop and Think: Sentiment Analysis

Sometimes, a text message might seem complimentary, but really is not. Is the customer happy if he tweets "Nice job, United. You only lost one of my bags this time." Think of some other short messages where sarcasm or slang makes sentiment analysis difficult. How would you write a product or service review that is unambiguously positive, negative, or neutral? How would you write a review whose sentiment is difficult to determine?

8.7 Key Points in Chapter Eight

- Classification is the systematic assignment of resources to a system of intentional categories, often institutional ones.

  (See §8.1 Introduction (page 319))

- A classification system is foremost a specification for the logical arrangement of resources because there are usually many possible and often arbitrary mappings of logical locations to physical ones.

  (See §8.1.3 Classification vs. Physical Arrangement (page 322))

- A classification creates structure in the organizing system that increases the variety and capability of the interactions it can support.

(See §8.2.1.2 Classifications Support Interactions (page 328))

- Classifications are always biased by the purposes, experiences, professions, politics, values, and other characteristics and preferences of the people making them.

  (See §8.2.3 Classification Is Biased (page 335))

- Three types of bias in technical systems are pre-existing, technical, and emergent bias.

  (See §8.2.3 Classification Is Biased (page 335))

- Classification schemes in which all possible categories to which resources can be assigned are defined explicitly are called *enumerative.*

  (See §8.1.4 Classification Schemes (page 323))

- When multiple resource properties are considered in a fixed sequence, each property creates another level in the system of categories and the classification scheme is *hierarchical* or *taxonomic.*

  (See §8.1.4 Classification Schemes (page 323))

- Classification and standardization are not identical, but they are closely related. Some classifications become standards, and some standards define new classifications.

  (See §8.1.5 Classification and Standardization (page 325))

- A standard is a published specification that is developed and maintained by consensus of all the relevant stakeholders in some domain by following a defined and transparent process.

  (See §8.1.5.3 Specifications vs. Standards (page 326))

- Standard semantics are especially important in industries or markets that have significant network effects where the value of a product depends on the number of interoperable or compatible products.

  (See §8.1.5.2 Institutional Semantics (page 325))

- The principle of *literary warrant* holds that a classification must be based only on the specific resources that are being classified.

  (See §8.2.2.1 Principles Embodied in the Classification Scheme (page 331))

- The *uniqueness principle* means the categories in a classification scheme are mutually exclusive. Thus, when a logical concept is assigned to a particular category, it cannot simultaneously be assigned to another category.

  (See §8.2.2.2 Principles for Assigning Resources to Categories (page 332))

- The general solution to satisfying the uniqueness principle in library classifications when resources do not clearly fit in a single category is to invent and follow a detailed set of often-arbitrary rules.

(See §8.2.2.2 Principles for Assigning Resources to Categories (page 332))

- Categories sometimes change slowly, but they can also change quickly and radically as a result of technological, process, or geopolitical innovation or events.

  (See §8.2.2.3 Principles for Maintaining the Classification over Time (page 333))

- *Flexibility*, *extensibility*, and *hospitality* are synonyms for the degree to which the classification can accommodate new resources.

  (See §8.2.2.3 Principles for Maintaining the Classification over Time (page 333))

- Bibliographic classification is distinctive because of a legacy of physical arrangement and its scale and complexity.

  (See §8.3 Bibliographic Classification (page 338))

- *Faceted* classification systems enumerate all the categories needed to assign resources appropriately, but instead of combining them in advance in a fixed hierarchy, they are applied only if they are needed to sort resources with a particular combination of properties.

  (See §8.4 Faceted Classification (page 342))

- Facets should be independent dimensions, so a resource can have values of all of them while only having one value on each of them.

  (See §8.4.4.2 Design Principles and Pragmatics (page 351))

- Facet values must accommodate potential additions to the set of instances. Including an "Other" value is an easy way to ensure that a facet is flexible and hospitable to new instances, but it not desirable if all new instances will be assigned that value.

  (See §8.4.4.2 Design Principles and Pragmatics (page 351))

- Most tagging seems insufficiently principled to be considered classification, except when tags are treated as category labels or when decisions that make tagging more systematic turn a set of tags into a *tagsonomy*.

  (See §8.1.2 Classification vs. Tagging (page 321))

- A task or activity-based classification system is called a *taskonomy*.

  (See §8.5 Classification by Activity Structure (page 352))

- *Supervised* learning techniques start with a designed classification scheme and then train computers to assign new resources to the categories.

  (See §8.6 Computational Classification (page 353))