2. What Are Controlled Vocabularies?

A controlled vocabulary is an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain.

2.1. Purpose of Controlled Vocabularies

The purpose of controlled vocabularies is to organize information and to provide terminology to catalog and retrieve information. While capturing the richness of variant terms, controlled vocabularies also promote consistency in preferred terms and the assignment of the same terms to similar content.

Given that a shared goal of the cultural heritage community is to improve access to visual arts and material culture information, controlled vocabularies are essential. They are necessary at the indexing phase because without them catalogers will not consistently use the same term to refer to the same person, place, or thing. In the retrieval process, various end users may use different synonyms or more generic terms to refer to a given concept. End users are often not specialists and thus need to be guided because they may not know the correct term.

The most important functions of a controlled vocabulary are to gather together variant terms and synonyms for concepts and to link concepts in a logical order or sort them into categories. Are a *rose window* and a *Catherine wheel* the same thing? How is *pot-metal glass* related to the more general term *stained glass?* The links and relationships in a controlled vocabulary ensure that these connections are defined and maintained, for both cataloging and retrieval.

2.2. Display Information and Controlled Information

Records for cultural objects typically contain both descriptive data and administrative data, which are outlined and defined in *CCO* and *CDWA.* Data elements record an identification of the type of object, creation information, dates of creation, place of origin and current location, subject matter, and physical description, as well as administrative

information about provenance, history, acquisition, conservation, context related to other objects, and the published sources of this information.

Both descriptive and administrative data must be maintained in ways that will accommodate two categories of information: information intended for display to end users and information intended for retrieval. Information utilized for retrieval should be adapted for controlled vocabularies and controlled format.

Why are the display and indexing of information separate issues? Art and cultural heritage information provides unique challenges in display and retrieval. Information must be displayed to users in a way that allows expression of nuance, ambiguity, and uncertainty. The facts about cultural objects and their creators are not always known or straightforward, and it is misleading and contrary to the tenets of scholarship to fail to express this uncertainty. At the same time, efficient retrieval requires indexing according to consistent, well-defined rules and controlled terminology.

A successful catalog of art and cultural heritage information maintains a balance between flexible standards and consistent rules. On the one hand, it must be flexible in allowing the expression of uncertainty and ambiguity where the discipline requires it, while also accommodating nuance and differences in style between departments and institutions. On the other hand, it must apply rules consistently where it is most critical—namely, for information that is indexed for retrieval.

In the context of this book, the controlled fields in a record are specially formatted and often linked to controlled vocabularies (authorities), controlled lists, or ruled by formatting restrictions (e.g., formatting of numbers) to allow for successful retrieval.

For a full list of fields for art information and their requirements for free-text, controlled format, or controlled vocabulary, see *CDWA* (fields and rules) and *CCO* (detailed rules for a subset of the *CDWA* categories).

2.2.1. Display Information with Controlled Vocabularies

It is often necessary to allow fuzziness in the expression of information that at the same time must be retrievable via terminology from a controlled vocabulary; in certain key areas of a work record, this is accomplished by including separate display and indexing fields for the same information. For example, in the creation statement and in technique, medium, and support statements, the information may be complex and may include indications of uncertainty through the use of words such as *or* or *probably*.

The most effective way to express the nuances of such information is to use natural language in a display field and to index the same

information separately, using controlled vocabulary (typically contained in an authority file). In the examples below, the creator's role is indexed with controlled terms, and the identity of the creator is indexed as well. The Creator Description field is free text, and authority files control the other fields. See **Chapter 6: Local Authorities** for a discussion of authority files and local authorities.

> **Creator Description:** Vincent van Gogh (Dutch, 1853–1890)
> **Role:** painter **Identity:** Gogh, Vincent van
>
> **Creator Description:** Marco Ricci (Venetian, 1676–1730), figures by Sebastiano Ricci (Venetian, 1659–1734)
> **Role:** painter **Extent:** landscape | architecture
> **Identity:** Ricci, Marco
> **Role:** painter **Extent:** figures **Identity:** Ricci, Sebastiano
>
> **Creator Description:** primary painter and calligrapher was Dai Xi (Chinese, 1801–1860), with additional inscriptions and colophons added by other officials; commissioned by Wu Zhongzhun
> **Roles:** painter | calligrapher **Identity:** Dai Xi
> **Role:** patron **Identity:** Wu Zhongzhun

2.2.2. Controlled Vocabularies vs. Controlled Format

While controlled vocabularies are organized sets of controlled terminology values (often with other information as well), the term *controlled format* refers to rules concerning the allowable data types and formatting of information. Fields may have controlled format in addition to being linked to controlled vocabulary, or the controlled format may exist in the absence of any finite controlled list of acceptable values.

Controlled format may govern the expression of Unicode or other characters in either a free-text field or in a field that is linked to a controlled vocabulary. Controlled format is also suitable for recording measurements, geographic coordinates, and other information in fields where numbers or codes are used. Restrictions may be placed on the field in order to regulate the number of digits allowed, the expression of decimals and negative numbers, and so on, ideally in compliance with ISO, NISO, or another appropriate standard where possible.

The examples below juxtapose a set of materials fields that use display and controlled vocabulary fields with a set of measurements fields. Fields such as Role and Material Name contain controlled vocabulary. However, in the measurements fields, the numbers in Value are indexed with controlled format but not controlled vocabulary.

Materials/Techniques Description: egg-tempera paint with tooled gold-leaf halos on panel
Role: medium **Material Name:** egg tempera | gold leaf
Role: support **Material Name:** wood panel
Technique Name: painting | gold tooling

Dimensions Description: comprises 10 panels; overall: 280 × 215 × 17 cm (110¼ × 84⅝ × 6¾ inches)
Extent: components
Value: 10 **Type:** count
Value: 280 **Unit:** cm **Type:** height
Value: 215 **Unit:** cm **Type:** width
Value: 17 **Unit:** cm **Type:** depth

Controlled format is also typically used for dates, such as the date of discovery or date of creation of an artwork. For such dates, controlled fields may be used in combination with a Display Date field.

The issues involved in recording data about dates illustrate the necessity of displaying information in a way that accurately expresses nuance and ambiguity to the end user, while at the same time formatting the dates consistently to allow retrieval. A free-text field for a Display Date can be used to express complex concepts and nuance, as in the examples below.

Display Creation Date: probably 1711
Display Creation Date: ca. 1910–ca. 1915
Display Creation Date: designed in the 1470s, constructed 1584–1627

The Display Date field should be combined with controlled Earliest Date and Latest Date fields that contain beginning and ending limits to enable searches on spans of time. The cataloger may estimate Earliest and Latest dates to allow for the leeway required by expressions such as *ca., before,* or *probably.*

The controlled Earliest and Latest fields do not contain controlled vocabulary per se, but they require a controlled format in which only numbers are allowed. A minus sign can be used to express dates BCE (Before Current Era) as negative numbers; dates CE (Current Era) are positive numbers. A rule should be in place ensuring that the latest date is always greater than or equal to the earliest date.

Display Creation Date: ca. 1913
Earliest: 1908 **Latest:** 1918

Display Creation Date: constructed 286–199 BCE
Earliest:–286 **Latest:**–199

Display Creation Date: 12th century
Earliest: 1100 **Latest:** 1199

Display Creation Date: Middle Minoan Palace period, ca. 1600 BCE
Earliest:–1630 **Latest:**–1570

Display Creation Date: 1039 anno Hegirae (1630 CE)
Earliest: 1630 **Latest:** 1630

Date fields should typically be controlled through locally defined rules rather than with default rules contained in the system. Although most systems promote the use of a special data type called *date* with predefined rules, this standard date data type does not generally work because art information requires the expression of dates up to many thousands of years BCE, and standard date data types are intended only for more modern dates (e.g., allowing 8-byte integers that represent dates ranging from 1 January of the year 0001 through 31 December of the year 9999).

2.3. Types of Controlled Vocabularies

Most controlled vocabularies discussed in this book are structured vocabularies. A structured vocabulary emphasizes relationships between and among the concepts represented by the terms or names in a vocabulary.

2.3.1. Relationships in General

In the context of this book, the term *relationship* means a state of connectedness or an association between two things in a database—in this case, fields or tables in a database for a controlled vocabulary.

One important type of relationship is between equivalents; for example, *Harlem Renaissance* and *New Negro Renaissance* refer to the same cultural movement that flourished in New York City in the 1920s.

Other relationships in a structured vocabulary include links that organize terms and provide context; for example, when discussing architectural drawings, an *orthographic projection* is a type of (child of) *parallel projection* and a sibling of *axonometric projection,* all of which are organized under *processes and techniques.*

The most common types of controlled vocabularies used for art and architecture include subject heading lists, simple controlled lists, synonym ring lists, taxonomies, and thesauri. Many of the definitions

Fig. 2. Display fields, as illustrated in the tombstone for this painting, are often indexed. The Display Materials field is indexed with controlled vocabulary. The Display Measurements field is indexed with controlled format for the numbers and a controlled list for the unit (centimeters, millimeters, inches, feet, square feet, among others) and type (height, width, depth, weight, area, circumference, among others).

Bartolomeo Vivarini (Italian, active from ca. 1440, died after 1500); *Polyptych with Saint James Major, Madonna and Child, and Saints;* 1490; tempera and gold leaf on panel; comprises 10 panels; overall: 280 × 215 cm (110¼ × 84⅝ inches); J. Paul Getty Museum (Los Angeles, California); 71.PB.30.

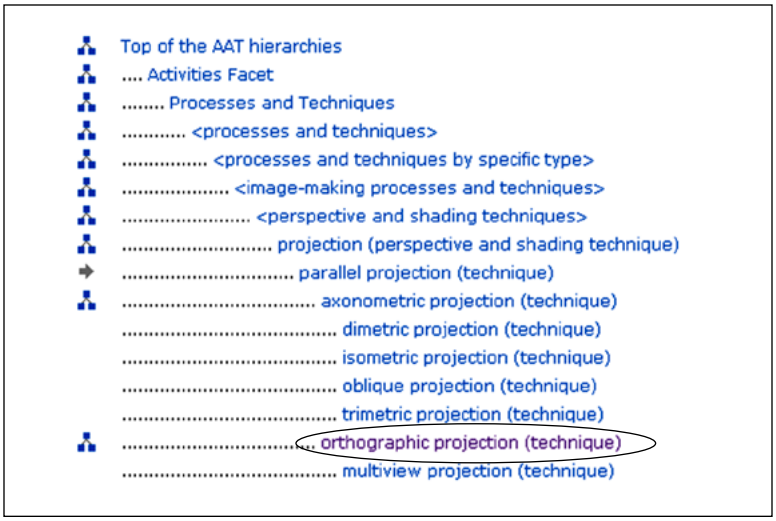

Fig. 3. A hierarchical display in the *AAT* illustrating *orthographic projection* with siblings and parents.

below are based on the discussions in *ANSI/NISO Z39.19-2005: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* and the related international standard, *ISO 2788:1986: Documentation—Guidelines for the Establishment and Development of Monolingual Thesauri.* Note that the types of vocabularies described below are not always mutually exclusive; for example, a single vocabulary can be both a thesaurus and an authority.

2.3.2. Subject Heading Lists

Subject headings, or simply *headings,* are uniform words or phrases intended to be assigned to books, articles, or other documents in order to describe the subject or topic of the texts and to group them with texts having similar subjects. The most commonly used subject headings in libraries in the United States are the *Library of Congress Subject Headings* (*LCSH*), which form a comprehensive list of preferred terms or strings, often with cross-references. Another well-known set of subject headings is the *Medical Subject Headings* (*MeSH*), which is used for indexing journal articles and books on medical science. *MeSH* incorporates a thesaurus structure with subject headings.

Subject heading lists are typically arranged in alphabetical order, with cross-references between the preferred, nonpreferred, and other related headings. This emphasis on a preferred entry and links

to synonyms may be found in other types of authorities. However, subject headings differ from the other vocabularies discussed here in the following fundamental way: precoordination of terminology is a characteristic of subject headings in that they combine several unique concepts together in a string. For example, the heading *Medieval bronze vessels* combines a period, a material, and a work type in one heading.

Subject heading lists typically include separate listings of standardized subheadings (e.g., geographic locations) that may be combined with designated headings according to prescribed rules. Various styles of subject heading displays are included in the examples below. *LCSH* displays two dashes and parentheses or periods as required, while other styles may omit punctuation or use colons or em dashes for compound phrases. In *LCSH, MeSH,* and other authorities, the parts of a compound heading may be stored in separate MARC format subfields to allow variations in displays as desired.

Bicycle racing--United States
Cat family (Mammals)--Literary collections
South Africa. Arts and Culture Task Group
Architecture—Ancient Egypt
Film history: Movements and styles
Embryonic and Fetal Development
Medieval bronze vessels
Great Britain Description and travel 1801–1900

2.3.2.1. Other Headings

Other types of headings or labels may be used to uniquely identify or disambiguate one vocabulary entry from another. That is, the vocabulary record itself represents a single unique person, place, or thing, but its name is displayed with information in addition to the name. For example, the name of a creator may be listed with a short biographical string (e.g., *Flemish painter, 1423–1549*) to form a heading or label for display in a work record. This type of heading or label is discussed in **Chapter 7: Constructing a Vocabulary or Authority.**

2.3.3. Controlled Lists

A controlled list is a simple list of terms used to control terminology. In a well-constructed controlled list, the following is true: each term is unique; terms are not overlapping in meaning; terms are all members of the same class (i.e., having the same level of rank in a classification system); terms are equal in granularity or specificity; and terms are arranged alphabetically or in another logical order. These lists are

also called *flat term lists* or *pick lists,* referring to the typical method of their implementation in an information system. Where appropriate, controlled lists should be derived from larger published standard vocabularies.

Controlled lists are usually designed for a very specific database or situation and may not have utility outside that context. They are best employed in certain fields of a database where a short list of values is appropriate and where terms are unlikely to have synonyms or ancillary information. However, as with any vocabulary for cataloging, it is preferable that definitions of the terms be made available to ensure consistency among catalogers. Below is an example of a controlled list for the Classification field in a work record.

architecture	manuscripts
armor	miscellaneous
books	paintings
coins	photographs
decorative arts	sculpture
drawings	site installation
implements	texts
jewelry	vessels

The advantage of such lists is that the cataloger or indexer has only a short list of terms from which to choose, thus ensuring more consistency and reducing the likelihood of error. In addition to the Classification field, examples of other art information fields that may benefit from a simple controlled list are Title Type (e.g., *artist's, descriptive, inscribed,* etc.), Title Language (e.g., *English, French, German, Italian, Spanish,* etc.), or Title Preference (e.g., *preferred, alternate*). Dozens of areas of a work record may be better suited for a short controlled list rather than a more complex controlled vocabulary. From the end-user perspective, such short lists may be easier to navigate than more complex lists, particularly for nonspecialist users.

2.3.4. Synonym Ring Lists

A synonym ring is a simple set of terms that are considered equivalent for the purpose of retrieval. Equivalence relationships in most controlled vocabularies should be made only between terms and names that have genuine synonymy or identical meanings. However, synonym rings are different. Even though they are classified as controlled vocabularies, they are almost always used in retrieval rather than indexing. They are used specifically to broaden retrieval (this is often referred to as query expansion): thus, synonym rings may in fact contain near-synonyms that have

similar or related meanings, rather than restricting themselves to only terms with true synonymy.

Typically, synonym rings occur as sets of flat lists and are used behind the scenes of an electronic information system. They are most useful for providing access to content that is represented in texts and other instances of natural, uncontrolled language.

Even though catalogers do not use synonym rings for indexing, subject experts should be involved in the creation of synonym rings for retrieval. The most successful synonym rings are constructed manually by subject matter experts who are also familiar with the specific content of the information system, user expectations, and likely searches.

In the example below, synonym rings (each represented in an individual row) represent true synonyms as well as more generic terms and other terms that are related within the specific context of a given text. The example could represent a partial synonym ring list for a text about art depicting certain migrating birds. If a user enters *crows,* the search mechanism returns any text containing *birds* or any of the other terms in the same synonym ring as *crows.* Even though these terms are not synonyms, the implementer has judged that these links make sense for broad retrieval in this particular text. Other automated retrieval strategies may be in place as well; for example, the search algorithm may automatically truncate the *s* to allow matches in English on both singular and plural forms.

birds, avian, storks, crows, ravens, herons, Ciconiidae, Corvus, Ardeidae
migration, nonmigratory, migratory, travel, flying, altitude
clouds, cumulus, nimbus, storm clouds, cloudy
wind, windy, windstorm, wind damage, air flow, jet stream

2.3.5. Authority Files

An authority file is a set of established names or headings and cross-references to the preferred form from variant or alternate forms. Illustrated on the following page is the *LCNAF*—the *Library of Congress/NACO (Name Authority Cooperative Program) Authority File*—an authority widely used in libraries in North America.

Common types of authority files are name authority files and subject heading authority files. However, any listing of terms, names, or headings that distinguishes between a preferred term, name, or heading and alternate or variant names may be used as an authority. In other words, almost any type of controlled vocabulary—with the exception of a synonym ring list—may be used as an authority.

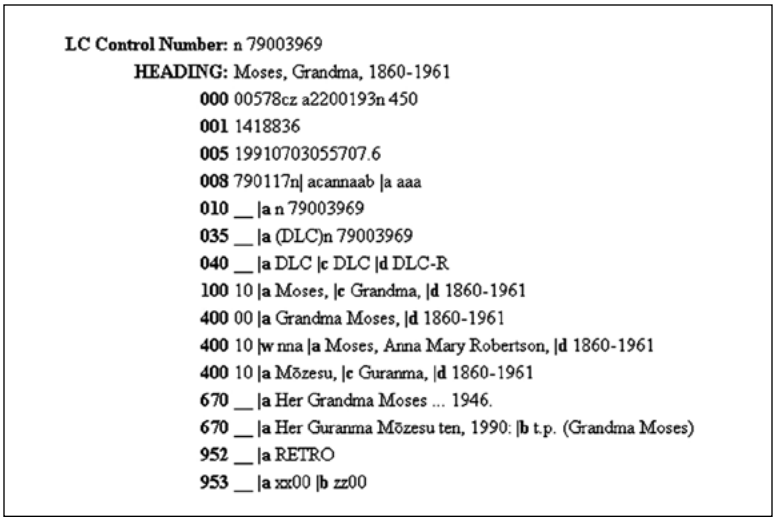

Fig. 4. The *LCNAF* record for *Grandma Moses*, illustrating the established heading and cross-references for this artist.

Authority control refers as much to the methodology as to a particular controlled vocabulary. If a controlled vocabulary is accepted by a given community as authoritative, and if it is used in order to provide consistency in data, it is being used as an authority. A local authority file is often compiled from terminology from one or more published standard controlled vocabularies. The establishment of local authorities is discussed in **Chapter 6: Local Authorities.**

2.3.6. Taxonomies

A taxonomy is an orderly classification for a defined domain. It may also be known as a *faceted vocabulary.* It comprises controlled vocabulary terms (generally only preferred terms) organized into a hierarchical structure. Each term in a taxonomy is in one or more parent/child (broader/narrower) relationships to other terms in the taxonomy. There can be different types of parent/child relationships, such as whole/part, genus/species, or instance relationships. However, in good practice, all children of a given parent share the same type of relationship.

A taxonomy may differ from a thesaurus in that it generally has shallower hierarchies and a less complicated structure. For example, it often has no equivalent (synonyms or variant terms) or related terms (associative relationships). The scientific classifications of animals and plants are well-known examples of taxonomies. A partial display of

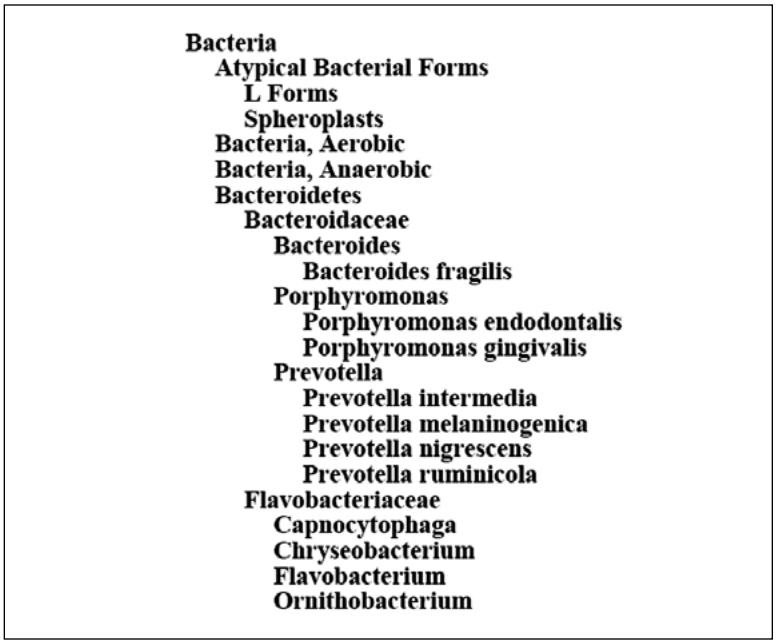

Fig. 5. A display of data from the U.S. National Center for Biotechnology Information illustrating the taxonomic placement of *Flavobacteriaceae* with siblings and broader and narrower contexts.

Flavobacteria in the taxonomy of the U.S. National Center for Biotechnology Information is above.

In common usage, the term *taxonomy* may also refer to any classification or placement of terms or headings into categories, particularly a controlled vocabulary used as a navigation structure for a Web site.

2.3.7. Alphanumeric Classification Schemes

Alphanumeric classification schemes are controlled codes (letters or numbers, or both letters and numbers) that represent concepts or headings. They generally have an implied taxonomy that can be surmised from the codes. The Dewey Decimal Classification (DDC) system is an example of a numeric classification scheme with which many people are familiar, given that it is one of the two major systems used in libraries in the United States (the other is the Library of Congress Classification [LCC] system). In the Dewey system, the universe of knowledge is divided into sets of three-digit numbers. The arts are represented in the 700-number series; sculpture is represented by numbers between 730 and 739. For example, the number 735 has been established to indicate

sculpture after the year 1400 CE. To that number may be added additional decimal indicators to further specify the topic by geographic or other categories. For example, 735.942 refers to sculpture dating after 1400 in England, because the extension 9 indicates geographic area, 4 indicates Europe, and 2 indicates England.

An alphanumeric classification scheme used for the iconography of art is *Iconclass,* discussed in **Chapter 4: Vocabularies for Cultural Objects.**

2.3.8. Thesauri

A thesaurus combines the characteristics of synonym ring lists and taxonomies, together with additional features. A thesaurus is a semantic network of unique concepts, including relationships between synonyms, broader and narrower (parent/child) contexts, and other related concepts. Thesauri may be monolingual or multilingual. Thesauri may contain three types of relationships: equivalence (synonym), hierarchical (whole/part, genus/species, or instance), and associative.

Thesauri may also include additional peripheral or explanatory information about a concept, including a definition (or scope note), bibliographic citations, and so on. A thesaurus is more complex than a simple list, synonym ring list, or simple taxonomy. Thesauri employ the versatile and powerful vocabulary control generally recommended for use as authorities in databases relating to art and cultural heritage.

The primary type of vocabulary discussed in this book is a thesaurus. Thesauri that contain art terminology include the Getty vocabularies, Chenhall's *Nomenclature,* and the *TGM,* which are discussed in **Chapter 4: Vocabularies for Cultural Objects.**

The term *thesaurus* may also be used for any controlled vocabulary arranged in a known order, displayed with standardized relationship indicators, and generally used for browsing in postcoordinated information storage and retrieval systems.

2.3.9. Ontologies

Whereas the vocabularies discussed above are the ones most commonly used for art information, discussions of controlled vocabularies may also include ontologies.

In common usage in computer science, an ontology is a formal, machine-readable specification of a conceptual model in which concepts, properties, relationships, functions, constraints, and axioms are all explicitly defined. Such an ontology is not a controlled vocabulary, but it uses one or more controlled vocabularies for a defined domain and expresses the vocabulary in a representative language that has a grammar

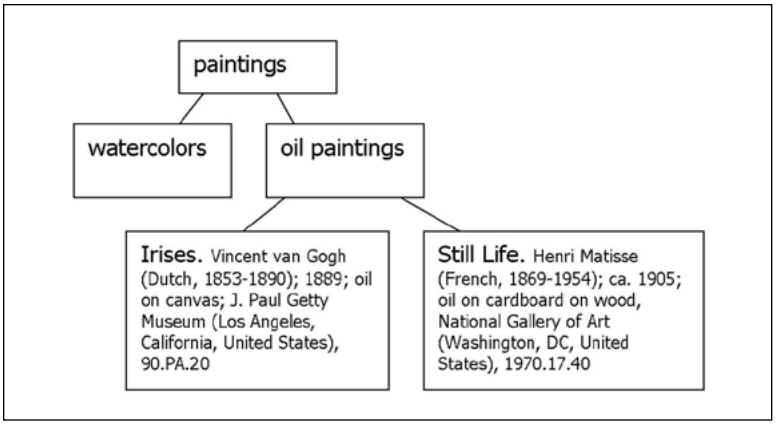

Fig. 6. A detail of a sample ontology for Vincent van Gogh's *Irises* and Henri Matisse's *Still Life,* illustrating how the works are part of a subset of *oil paintings* under the category *paintings.*

for using vocabulary terms to express something meaningful. Ontologies generally divide the realm of knowledge that they represent into the following areas: individuals, classes, attributes, relations, and events. The grammar of the ontology links these areas together by formal constraints that determine how the vocabulary terms or phrases may be used together. There are several grammars or languages for ontologies, both proprietary and standards-based. An ontology is used to make queries and assertions.

Ontologies have some characteristics in common with faceted taxonomies and thesauri, but ontologies use strict semantic relationships among terms and attributes with the goal of knowledge representation in machine-readable form, whereas thesauri provide tools for cataloging and retrieval.

Ontologies are used in the Semantic Web, artificial intelligence, software engineering, and information architecture as a form of knowledge representation in electronic form about a particular domain of knowledge.

In the example above, each item in the ontology belongs to the subclass above it. Items can also belong to various other classes, although the relationships may be different. For example, a watercolor is a painting, but it may also be classified as a drawing because it is a work on paper. Van Gogh's *Irises* could be classified with oil paintings (with the relationship type *medium is*) but also with Post-Impressionist art (with relationship type *style/period is*). Relationships in ontologies are defined according to strict rules, which are different than the equivalence,

hierarchical, and associative relationships used for thesauri and other vocabularies discussed in this book.

2.3.10. Folksonomies

Folksonomy is a neologism referring to an assemblage of concepts represented by terms and names (called *tags*) that are compiled through social tagging. *Social tagging* is the decentralized practice and method by which individuals and groups create, manage, and share tags (terms, names, etc.) to annotate and categorize digital resources in an online social environment. This method is also referred to as *social classification, social indexing, mob indexing,* and *folk categorization.* Social tagging is not necessarily collaborative, because the effort is typically not organized; individuals are not actually working together or in concert, and standardization and common vocabulary are not employed.

Folksonomies do not typically have hierarchical structure or preferred terms for concepts, and they may not even cluster synonyms. They are not considered authoritative because they are typically not compiled by experts. Furthermore, they are by definition not applied to documents by professional indexers. Given that it is impossible for the large and varied community of creators and users of Web content to independently add metadata in a consistent manner, folksonomies are generally characterized by nonstandard, idiosyncratic terminology. Although they do not support organized searching and other types of browsing as well as tags from controlled vocabularies applied by professionals, folksonomies can be useful in situations where controlled tagging is not possible: they can also provide additional access points not included in more formal vocabularies. There may be great potential for enhanced retrieval by linking terms and names from folksonomies to more rigorously structured controlled vocabularies.