Defining Bibliographic 'Works:' Naïve classification for Terminology Generation

RICHARD P. SMIRAGLIA
Long Island University, E.U.A.

ABSTRACT: Empirical methods are effective for the generation of descriptive terminology, especially concerning new or little-understood phenomena. Direct observation provides a critical base point from which terminology can be generated for further rational analysis. This procedure has been called "naïve classification" because it closely parallels the process of category generation in the laboratory. In the present paper, the evolution of terminology used to describe the phenomena concerning bibliographic "works" is demonstrated as a form of naïve classification.

1. COGNITION, CLASSIFICATION, NAÏVE CLASSIFICATION

Classification is a near universal human phenomenon. When you say hello to a child and she says "Grandma," it is because she has recognized that you are her grandmother, and therefore not any other person. She has created a classification with at least two categories—grandmother and not-grandmother—and she has assigned you as a member of one category and therefore not a member of the other. It is simple cognition at one level, but it is also classification. Therefore classification permeates human activity.

In a formal sense classification is acknowledged to have two uses for scholarship. First, scientists use classification to order their phenomena of study—often called *taxonomy*, this activity is essential for the advancement of knowledge. The second major use of classification is for the ordering of useful knowledge, and this can be seen as activity that crosses a broad spectrum of uses from social to scholarly to bibliographic. Dahlberg (2006) points out the common methods in use, which are the designation of objects of interest (knowledge elements), designation of the conceptual parameters for categories and relationships among them (knowledge units; i.e., this is the building of ontology), and the mapping of entities to the designated structure (knowledge systems). Your doctor checks a box on a diagnostic form, you find the tomatoes in the Italian foods section of your supermarket, you find mystery next to biography at your public library—all of these are examples of the use of classification for the useful ordering of knowledge.

In this paper I am concerned with the former definition of classification—that for use in the discovery process of science. I am particularly eager to demonstrate how this process has been used over the past two decades to investigate the nature of the bibliographic phenomenon called "works," and the phenomenon of "instantiation" that is shared among diverse types of information objects. I bring this work to the attention of the present symposium as an example of the appropriate use of knowledge organization for the generation of terminology in the field of library and information science.

The specific research stream devoted to the definition of "works" and "instantiation" constitutes an example of what Beghtol (2003) has called "naïve classification." The direct observation of phenomena provides a critical base point from which rationalism may be applied to provide conceptual connections and enhance understanding of research phenomena. More simply put, the researcher investigates empirically, records his observations, names his categories, and then adjusts the categories and the terms over time as new data are discovered. In this way naïve classification of works has yielded a rich typology of instantiation.

The empirical derivation of knowledge-elements, particularly in developing or evolving KO systems, provides a basis upon which conceptual systems can be built. This is exactly Beghtol's process of naïve classification for use in evolving scholarship. According to Beghtol (2003, 66), the process of naïve classification has many uses, including the discovery of gaps in knowledge, the reconstruction of historical evidence, and the revision or amplification of existing knowledge organization schema, among others. The process requires the scholar to articulate the purpose for his work so as to limit the empirical parameters, and then a variety of techniques may be employed, including paradigm-identification, and ordering (hierarchy, tree-structure, faceting) techniques.

Beghtol refers to studies that report naïve classification of Chinese plates, paintings, religions, photographs, 13th century Spanish silks, and child-rearing practices, among others. Green and Fallgren (2007) use the techniques to analyze document structures for the revision of the *Dewey Decimal Classification*. Let us look at a very simple example. We begin by identifying the phenomena observed and creating simple groupings. Figure 1 shows a set of observations divided into two clusters.

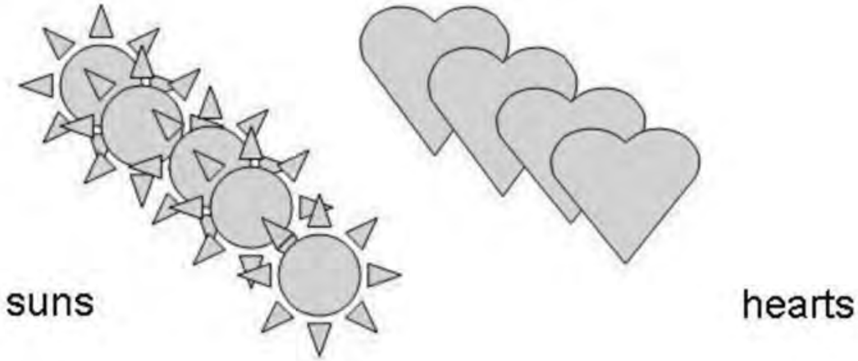

Figure 1. Suns and Hearts

In this illustration we have a naïve classification. There are nine objects in our laboratory, and they clearly can be divided by observable likeness into two categories, hearts and suns. There are five suns and four hearts. This classification is complete, because it identifies the knowledge elements (objects of study), knowledge units (hearts and suns), and a knowledge system "Suns and Hearts." The classification is naïve because it represents merely the grouping of observations in this particular instance. We do not yet know why the suns are not hearts or why the hearts are not suns, we do not know what the suns and hearts have in common except that they both are found in this observation, we do not know why there are more suns than hearts or fewer hearts than suns, and we do not know why there are not other objects, such as moons or lightening bolts or stars. In research, the use of naïve classification is intended for just this process—to identify the paradigm and to create a matrix of its contents from which hypotheses can be generated to structure future research.

In the remainder of this paper we will define the concepts of "works" and "instantiation," we will briefly summarize empirical methods that have been applied to their analysis and the naïve classifications that have resulted, and we will examine the typology of instantiation that exists at present. Finally we will look at this typology revealed by naïve classification over and against several rational schemes that have arisen elsewhere in literature. In this way I hope simply to demonstrate the usefulness of naïve classification for the advancement of knowledge through empirical research.

2. Works, and instantiation

In library and information science we are much concerned with the artifacts of recorded knowledge, and in particular, we are motivated to organize these artifacts for retrieval, either in whole (as when a reader seeks a particular book) or in part (as when a scholar seeks a particular fact). We sometimes speak of these artifacts in specific terms: books, maps, journals, scores, videos, etc.—and we sometimes speak of them in generic terms: documents. In either event, we make a clear distinction between the objects themselves and their content. A "work" is the intellectual content of a document. But a work is much more than that because research has shown that it is also the deliberate creation of intellectual content that is intended by its creator to be communicated to some audience. Thus a seashell may be described as a document because for some scholars it is considered informative, but it is not a work, per se. But a sculpture of a sea shell is a work, because its sculptor has intended it to be appreciated by its audience.

Most cataloging codes make a distinction between works and items, but until recently the distinction was not clear. Further, the structure of catalogs has failed to make organization and retrieval of works a primary objective. This becomes a problem when a single work exists in multiples—as when a novel becomes a screenplay, which becomes a motion picture, which becomes a video, whose music is issued on a CD, and so on. In these cases all of these diverse publications are likely to have the same name (e.g., *Harry Potter and the Order of the Phoenix*), therefore it becomes necessary to introduce order in the display to disambiguate the cluster. Another important aspect of the problem of works in the catalog is that of collocating all of the works of a particular creator. Again, ordering devices are required to disambiguate the collocated works under a particular heading (e.g. all Harry Potter books and their translations and all movies of Harry Potter books). These problems are treated fully in Smiraglia (2001).

The phenomenon of multiple iterations of a work over time is called "instantiation." Instantiation occurs when some cultural catalyst acts like a market force and causes a work that has entered a canon to be translated, adapted, edited and reissued over time. The longer the period of cultural catalysm the greater the number of instantiations, and the more recent (say, the current decade, rather than the 19th century) the greater the complexity of the instantiation set. That is, we know from prior research that significant works have many editions and translations, and that works that have become significant recently will have many multi-media iterations. The problem for research has been to move beyond the simple concept of works present in cataloging codes and to understand the phenomenon more fully. The origins of this phenomenon are treated in Smiraglia (2005b).

2.1 Empirical methods for studying works and instantiations

Research about works and instantiations has followed an empirical trajectory for more than two decades. Many studies (detailed in Smiraglia 2001, and treated as meta-analysis in Smiraglia 2002) have yielded interesting data about works and the extent of their instantiation. Roughly consistent statistical data have emerged showing that in most libraries the proportion of instantiating works follows roughly Lotka's law (thus around 30-40% instantiate), while most other works are unique singletons. These data emerge from empirical studies of library collections and bibliographic utilities in which a common methodology was employed. This methodology can be summarized briefly:

- Identify a sampling frame (e.g. an online catalog);
- Select a random sample of bibliographic records;
- Adjust the sample to make sure each "work" is unique in the sample, and to make certain each instantiated work in the sample is the progenitor in the sampling frame;
- Constitute the resulting list of works;
- Search all works in the bibliographic utilities and compile all sets of instantiations;
- Analyze the instantiation sets.

A taxonomy of instantiation was created by Smiraglia (1992), which included these categories:

1. Simultaneous derivations;
2. Successive derivations;
3. Translations;
4. Amplifications;
5. Extractions;
6. Adaptations; and,
7. Performances.

As research progressed new categories were discovered in the data. For instance, works that have large instantiation sets are often derived from earlier works, thus the category “predecessor work” was added. Also, many works are issued together with other works (a map, for instance, inside a book); these are called “accompanying works.” Vellucci (1997) found two categories of uniquely musical instantiation, which she called “notational transcription” and “musical presentation.” Smiraglia (2007) also found a category that occurs among best-selling works, which he called “persistent works.” In this way, the principle of naïve classification expanded the list of instantiation types for published works. As Beghtol suggests, naïve classification leads to new hypotheses, and in this case semiotic analysis of instantiation demonstrated that some of these categories represented new publications in which there had been little change in semantic content, while others represented tremendous change in semantic content. Thus, two meta-categories were developed—derivation, to describe instantiation with little or no change, and mutation, to describe instantiation with much change—that can be applied roughly hierarchically, thus:

Table 1. Naïve classification of instantiation types for publications

Derivations
simultaneous editions
successive editions
predecessors
amplificacions
extractions
accompanying materials
musical presentations
notational transcription
persistent works
Mutations
translations
adaptations
performances

2.2 CASE STUDY FOR STUDYING INSTANTIATION AMONG ARTIFACTS AND ARCHIVES

To extend this research, study of museum artifacts and archival documents was undertaken. Although the concept of "works" remains limited to deliberately created intellectual entities, the question was whether other informing objects would have instantiation networks that might require disambiguation for information retrieval. That is, while our naturally occurring sea shell is not a work, per se, if it is collected by a repository, there will be representations of it in the repository. Is it possible these would undergo the phenomenon of instantiation? The answer was resoundingly affirmative. Analysis of eight Etruscan artifacts from the University of Pennsylvania Museum of Archaeology and Anthropology was undertaken using case study method. The results are reported fully in Smiraglia (2005a). In this case, for each artifact the museum's archives were searched thoroughly to locate all representations in-house. Subsequently, published literature was searched to locate additional representations. The results yielded a typology of both representations (photographs, models, etc.) and metadata sets (narrative descriptions), and both were discovered to exist in multiple instantiations both in-house and in publication. Further, additional research was conducted on the collected archival papers of the United States Merchant Marine Academy's Class of 1942 Archives (reported in Smiraglia 2006). The papers were analyzed prior to digitization, and again instantiation among representations turned up. In this case we found correspondence that was present as typescript, carbon copy, and photocopy, we found photographs that reappeared on post-cards, which were in turn scanned and used as digital images on computer-generated documents, and so on.

The resulting naïve classification of works and instantiation can be summarized in its present form in the following table (Table 2).

Table 2. Naïve Classification of Comparative Instantiation Typologies

Bibliographic works	Artifacts - Metadata	Artifacts - Representations	Personal Papers
simultaneous editions	finding aids	field photos	Photocopies

▶

successive editions	field notes	working images	Carbon copies
amplifications	letters	exhibition color images	Photos
extractions	conservation treatment notes	digitized exhibition images	postcard with photo
Musical presentations	register descriptions; object cards	conservation photos	digitized scan of postcard with photo
National Transcriptions	image order invoices	archived photographic negatives	reprint of photo
Persistent works	museum database records	archived photographic prints	digitized scan of photo
	catalog card records	archived photographic transparences	
translations		object reproductions	
adaptations		drawings	
performances		3D models	
predecessors			
Accompanying materials			

From left to right are seen the typologies of instantiation that arose from the research. It is apparent that the bibliographic typology indicates types of publications, whereas the typologies from the museum and the archives indicate document or artifact types. There is no horizontal correspondence to be drawn from this table; rather the entries are simply arrayed by type for display. But the hierarchical grouping of derivations and mutations that we saw in table 1 persists here. The terms entered below the solid line represent instantiation types that could be considered to be mutations, because the content of the original appears in altered form. All of the instantiation types above the line are derivations—they represent re-iterations of the original content.

3. The value of naïve classification for terminology generation

The purpose of this paper was to demonstrate the use of empirical means to generate a KO structure, and in particular to highlight the evolution from a naïve classification of simple knowledge-elements to a more useful model of a phenomenon, incorporating conceptual knowledge-units. Following Dahlberg (2006) we have identified a set of typologies of instantiating information objects as knowledge elements, we have grouped them into knowledge units, and we have arrayed them hierarchically into a knowledge system. Following Beghtol (2003) we have used the paradigm-generating approach of naïve classification by declaring our purpose (to study works and instantiations for the purpose of aiding disambiguation in information retrieval), we have identified gaps in knowledge of the phenomenon and generated new hypotheses sequentially, and as the research progressed we have allowed the classification to evolve.

In this instance what began as empirical research has yielded a base level naïve classification, from which additional future hypotheses can be generated. We have not created a sophisticated taxonomy of instantiation. Rather, we have generated a typology of derivations and mutations of informing objects. This naïve classification then, suggests the value of continued research into the phenomenon of instantiation. It demonstrates the potential breadth of terms that should be incorporated in any future taxonomy. And it suggests that scholars of knowledge organization in library and information science might well continue to benefit from the use of empirical methods for terminology generation.

References

Beghtol, C. (2003). Classification for information retrieval and classification for knowledge discovery: Relationships between "professional" and "naïve" classifications. *Knowledge organization* 30: 64-73.

Dahlberg, I. (2006). Knowledge organization: a new science? *Knowledge organization* 33: 11-19.

Green, Rebecca and Fallgren, Nancy. 2007. Anticipating new media: a faceted classification of material types. In Tennis, J. ed. *North American Symposium on Knowledge Organization* http://dlist.sir.arizon.edu/1911.

Smiraglia, Richard P. 1992. Authority control and the extent of derivative bibliographic relationships. PhD. Dissertation. University of Chicago.

Smiraglia, Richard P. 2001. *The nature of a work: implications for the organization of knowledge*. Lanham, MD: Scarecrow.

Smiraglia, Richard P. 2002. Further progress in theory in knowledge organization. *Canadian journal of information and library science* 26 n2/3: 30-49.

Smiraglia, Richard P. 2005a. "Content metadata—an analysis of Etruscan artifactsin a museum of archeology." *Cataloging & classification quarterly* 40 n3/4:135-51.

Smiraglia, Richard P. 2005b. Instantiation: Toward a theory. In Vaughan, L., ed., *Data, information, and knowledge in a networked world;* Annual conference of the Canadian Association for Information Science … London, Ontario, June 2-4 2005. Available http://www.cais-acsi.ca/2005proceedings.htm.

Smiraglia, R.P. (2006). Empiricism as the basis for metadata categorization: expanding the case for instantiation with archival documents. In Budin, G., Swertz, C. and Mitgutsch, K., eds., *Knowledge organization and the global learning society;* Proceedings of the 9th ISKO International Conference, Vienna, July 4-7 2006, pp. 383-88.

Smiraglia, Richard P. 2007. The 'works' phenomenon and best selling books. *Cataloging & classification quarterly* 44n3/4: 179-95..

Vellucci, Sherry L. 1997. *Bibliographic relationships in music catalogs*. Lanham, Md.: Scarecrow Press.