Empiricism as the Basis for Metadata Categorisation: Expanding the Case for Instantiation with Archival Documents

Abstract: Metadata schemas tend to be rationally ordered instruments for the categorization of data about information objects. Instantiation has been demonstrated to be a universal phenomenon. Empirical analysis, both positivist and qualitative, has contributed to typologies of the properties of instantiation. This yields a naïve knowledge organization schema of instantiation. Bibliographic, museum, and archival analyses are compared to demonstrate the value of empirical derivation of categories. In this instance categories, once derived, are demonstrated to represent properties yielding typologies. The empirical generation of categories for knowledge organization is demonstrated.

1. Introduction
In the sphere of metadata, and in particular in the area of knowledge representation, the emerging concept of instantiation holds promise for the construction of increasingly sophisticated retrieval mechanisms. Instantiation” is the phenomenon addressed by research into bibliographic ‘works,’ and more recently ‘content genealogy’ of artifactual representations. Specifically, an instantiation of a work exists whenever the work is manifest in physical form (in a book, for example). A problem arises when multiple instantiations of a work (several editions, translations, etc.) exist and must be collocated (i.e., classified alphabetically) in a retrieval system with sufficient information to assist in the selection of the instantiation of interest to a searcher. Similarly, unique artifacts can be represented by metadata or images (called representations), which can exist in multiple instantiations (a photographic negative, a print, its digital descendent, etc.). The same is true of the representations of archival documents, which might exist in paper photocopies, digital images, and so forth.

Metadata are categorical descriptors of information resources, which often are used as alphabetico-classed segments of thesaur-faceted strings for information retrieval. Multiple instantiations of a work (editions, translations, etc.) that appear in an information retrieval system must be collocated (i.e., caused to appear to be adjacent). In such a system, there must be sufficient information to assist in the selection of the instantiation of interest to a searcher. So far, in bibliographic retrieval systems, collocation of this type has been achieved by the use of alphabetical classifiers called uniform titles.

Metadata schemas are generated rationally and sometimes pragmatically, but generally without evidentiary source. In this paper, a sequence of research projects is summarized to demonstrate the value of empirical observation in the creation of metadata for ordering collocated instantiations. Studies of bibliographic works, museum artifacts, and personal papers are used to generate typologies of instantiation.

2. The Role of Empiricism
The use of empirical research methods in KO is rare, and has been confined largely to the positivist testing of components of retrieval systems. There has been little testing of principles or assumptions that form the basis of bibliographic control practice; such testing
could generate sufficient evidence to turn principles into theories. This absence of empirical research (or other appropriate methodological approaches) in Knowledge Organization (KO) was criticized by Hjørland (2003), who suggested developments in practice were likely more technology-driven than theoretically justifiable. Through historical and epistemological analysis of KO he argued for the formal use of four fundamental methods, which parallel four fundamental epistemic stances (2003, 107): empiricism (observations and inductions), rationalism (principles of pure reason, deductions), historicism (study of context and development; explicating pre-understanding), and pragmatism (analysis of goals, values and consequences in both subject and object).

That said, a major contribution to theory has come from research into the content and formation of components of library catalogs. In particular, several projects generated substantial evidence about the structure of catalogs and the origins of their source data. One result has been increased confidence in the generalization due to increased confidence in external validity. These projects—Taylor’s investigation of name headings (Taylor 1982), Potter’s application of Lotka’s law (Potter 1980), Tillett’s analysis of bibliographic relationships (Tillett 1987), and the subsequent suite of research by this author into the ‘works’ phenomenon—are presented in meta-analytical form in Smiraglia (2002).

Hjørland’s appeal to activity theory (2003, 98) is probably most directly relevant to the present project. The act of naming objects (documents, artifacts, records, and their content, to be specific) is the action of facilitating their use. Terminology thus used cannot be neutral, because of the influence it brings to the activity of facilitating (or obfuscating) use. Metadata schemas tend to categorize based on rationally deduced categorizations of objects and their component parts. Thus rationalized, the schema predetermine the potential use of intellectual content by facilitating or limiting its retrieval. What I mean to suggest is that the base point for construction of metadata schema, particularly those designed to embrace intellectual content, should be empirical observation of the content itself. More colloquially put, letting the documents, artifacts, and records speak for themselves allows the creators (all of them, from authors to printers) to play a role in the use of the intellectual content.

In an analysis of metadata schema and their evolution, Greenberg (2005, 30) emphasized the importance of empirical research, specifically into the problem of instantiation with the purpose of informing data modeling techniques for object representation. Data modeling, she suggested, is a way of typing objects—that is, naming them to facilitate their use. When multiple instantiations exist side by side in a retrieval context, explicit, content-driven typologies are required to sort instantiations. Empirical observation of the instantiations in a variety of networks can yield inherent typology, such as those that are demonstrated below.

3. Empirical Research and Instantiation

Studies to date have demonstrated some consistent theoretical parameters for the concept of instantiation, even across bibliographic and artifactual borders. In the bibliographic analysis of works, samples drawn from online catalogs formed the basis of study. For each work, all instantiations extant in the bibliographic networks OCLC and RLIN were identified and sorted according to a taxonomy. The initial taxonomy, created for analysis of instantiations of works from an academic library catalog, included seven categories: simultaneous editions; successive editions; translations; amplifications; extractions; adaptations; and, performances (Smiraglia 2001, 42). Subsequent research using a sample of works drawn from the OCLC WorldCat (a union catalog), yielded two new types. One was ‘predecessor,’ which was used for notes or sketches of the work under study, as well as for instances such as novels or screenplays developed from short stories. The other was ‘accompanying material,’ which was used when the work was included in more than one
medium, such as a textbook accompanied by its text on a computer disk. Vellucci studied musical works and generated two more types of instantiation that were music-specific—‘musical presentation,’ and ‘notational transcription.’ More recently Smiraglia (forthcoming) studied best-selling books of the twentieth century, and reported the presence of a new category—‘persistent works’—used to denote works whose instantiation networks develop after the initial publishing frenzy that usually accompanies best sellers.

Interestingly, continued tabulation of instantiation networks revealed that the terms represented not so many categories as properties. The categories were not mutually exclusive, but could appear together in the same instantiation. That is, just as a male may be tall, so a translation may have a successive edition or appear in a commentary or both. This suggests that what we are dealing with is not so much taxonomy as typology.

To extend the concept of instantiation, a set of Etruscan artifacts from the University of Pennsylvania Museum of Archeology and Anthropology was examined. The artifacts were, of course, unique. But for each, many representations (ranging from photographs to models), as well as metadata descriptions, existed both in the museum and in publications. It turned out that the representations were often multiply instantiated (Smiraglia 2005a). The instantiations of artifacts, like those of bibliographic works, yielded a typology as well. The typology has two sets. For metadata: finding aids; field notes; letters; conservation treatment notes; register descriptions or object cards; image order invoices; museum database records; and catalog card records. For artifactual representations: field photos; drawings; working images; 3d models; exhibition color images, and digitized images; conservation photos; photo archives, including negatives, prints, and transparencies; and object reproductions. Every representation, metadata or artifactual, was accompanied by an object description, whose components constitute an object entity: object type, material, culture, source, collector, and date acquired.

A natural question, then, is: can the concept of instantiation be extended to unprocessed, raw data, as in the case of archival evidence? The extension to museum artifacts was fruitful, and knowledge of the digitization of archival finding aids, as well as the impending digitization of archival resources meant that it was likely instantiation could be observed in archival contexts.

The U.S. Merchant Marine Academy at Kings Point, Long Island, New York, houses an extraordinary library with rich archives of former midshipmen. In particular, the Class of 1942 is a rich source of documentation. In particular, the Class of 1942 is a rich source of documentation. The archives consists of fourteen “folders” created by members of the class, and created for the purpose of leaving an historical record of the class. These were young men who chose the seafaring life as a future career opportunity between the ages of 16 and 18. Once admitted to the academy they had to deal with all of the problems of college life (getting along, passing courses, and paying tuition). And yet these young men were destined for greatness. They entered the academy in 1938. Shortly after the Peal Harbor event of December 7, 1941 they found themselves commissioned as officers in the merchant marine. And they were soon in the thick of war. These archives document their lives.

For this study one midshipman’s documents were examined. The collection was rich with letters, envelopes, binders, photographs, ship’s deck-logs, time-sheets, scholarship applications, and so forth. Of particular interest was a canvas-bound ring-binder containing ‘orders’—papers ordering military personnel from place to place. The binder was heavy. There is a note inside indicating the binder had been issued specifically because it would sink in ocean water. In the event a ship was boarded, these binders were to be tossed overboard so as to be lost to enemy intelligence.
But the enterprising graduates of this class have managed to compile large instantiation networks of the documents pertaining to their service. In the file we located photocopies, carbon copies, digitized scans of postcards containing photographs, scans of photos, photos alongside digitized scans of them, and documents together with their carbon copies and digitized scans of the originals. Instantiation is present even in this case. These documents demonstrate the occurrence of instantiation among archival entities. The analysis of instantiation demonstrates not only the universality of multiplicity among informing objects, it also demonstrates the usefulness of empirical strategies for generating categories and their descriptions.

4. Conclusions

The typologies from the three studies can be placed side by side in tabular form for visual impact (see Table 1). The table lends support to Hjørlands idea that activity theory can proscribe the categorizing activity of knowledge organization. We denote categories so as to assign information objects spatial loci within the acknowledged schema. We usually generate these schema rationally, but without reference to the content that is to be so schematized. In the present study we have seen how empirical evaluation (both positivist and qualitative) has yielded useful typologies of instantiation.

<table>
<thead>
<tr>
<th>Bibliographic Works</th>
<th>Artifacts--Metadata</th>
<th>Artifacts--Representations</th>
<th>Personal Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>simultaneous editions</td>
<td>-finding aids</td>
<td>-field photos</td>
<td>Photocopies</td>
</tr>
<tr>
<td>successive editions</td>
<td>-field notes</td>
<td>-working images</td>
<td>Carbon copies</td>
</tr>
<tr>
<td>predecessors</td>
<td>-letters</td>
<td>-exhibition color images</td>
<td>Photos</td>
</tr>
<tr>
<td>amplifications</td>
<td>-conservation treatment notes</td>
<td>-digitized exhibition images</td>
<td>-postcard with photo</td>
</tr>
<tr>
<td>extractions</td>
<td>-register descriptions; object cards</td>
<td>-conservation photos</td>
<td>-digitized scan of postcard with photo</td>
</tr>
<tr>
<td>accompanying materials</td>
<td>-image order invoices</td>
<td>-archived photographic negatives</td>
<td>-reprint of photo</td>
</tr>
<tr>
<td>musical presentation</td>
<td>-museum database records</td>
<td>-archived photographic prints</td>
<td>-digitized scan of photo</td>
</tr>
<tr>
<td>notational transcription</td>
<td>-catalog card records</td>
<td>-archived photographic transparencies</td>
<td></td>
</tr>
<tr>
<td>persistent works</td>
<td>-finding aids</td>
<td></td>
<td></td>
</tr>
<tr>
<td>translations</td>
<td></td>
<td>-object reproductions</td>
<td></td>
</tr>
<tr>
<td>adaptations</td>
<td></td>
<td>-drawings</td>
<td></td>
</tr>
<tr>
<td>performances</td>
<td></td>
<td>-3D models</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Comparative Instantiation Typologies

The terms in these typologies, empirically derived, represent the properties of instantiation in three contexts, and yet they demonstrate the epistemological properties of mutation and derivation (see Smiraglia 2002b). Derivation denotes types or properties of instantiation in which intellectual content is unaltered; mutation denotes types or properties of instantiation in which intellectual content has been altered semantically or ideationally. In table 1, terms listed below the solid line represent mutations, which occur in both bibliographic and artifactual typologies. According to research to date, archival records and artifactual metadata typologies identify derivations.
In fact, these typologies represent a form of naïve classification system. As Beghtol suggests (2003, 66) the terms in the typologies discover and fill gaps in knowledge about instantiation, reconstruct empirically derived evidence, facilitate integration of findings, and suggest the complexity of the concept known formerly in knowledge organization and bibliographic control, simply, as ‘the work.’

Other consistent elements across these studies of instantiation include: 1) the concept of ‘canonicity’ as a predictor of instantiation; 2) the influence of time as a predictor of the degree of instantiation; 3) transitive relations demonstrated at points of change in semantic or ideational content as predictors of the type of instantiation; and 4) a continued association of the incidence of instantiation with Lotka’s law (Smiraglia 2005b).

These results demonstrate the importance of the phenomenon of instantiation for the design and implementation of information systems for a global learning society. Pan- and inter-institutional digital libraries incorporate representations of documentary, artifactual, and archival information resources. In all three cases, instantiation enriches the resource base, but threatens chaos in retrieval. Empirical derivation of instantiation typologies, as demonstrated here, suggests a realistic approach to metadata solutions.

6. References


Potter, William Gray. 1980. When names collide: Conflict in the catalog and AACR. Library resources & technical services 24: 3-16.


