Contents

Special Issue: Domain Ontologies, Part 1; guest editor Yejun Wu

Articles

L.P. Coladangelo.
Thesaurus and Ontology Construction for Contra Dance: Knowledge Organization of a North American Folk Dance Domain ................................................................. 523

Ann M. Graf.
Domain Analysis Applied to Online Graffiti Art Image Galleries to Reveal Knowledge Organization Structures Used Within an Outsider Art Community ....................... 543

Richard P. Smiraglia, Jihua Henry, Elizabeth Milonas, Chris Marchese and Sergey Zherebchevsky.
A Formal Taxonomy of Knowledge Organization: Meta-Analysis and Facet Analysis ................................................................. 558

B. Preedip Balaji and M. Dhanamjaya.
Towards a Model of Urban Studies Classification ........................................ 574

Jaime A. Pinto and Mauricio B. Almeida.
An applied ontology-Oriented Case Study to Distinguish Public and Private Institutions Through Their Documents ..................................................... 582

Gercina Ângela de Lima, Maria Luiza de Almeida Campos, Patrícia Lopes Ferreira França.

Xiaoguang Wang, Wanli Chang and Xu Tan.
Representing and Linking Dunhuang Cultural Heritage Information Resources Using Knowledge Graph ......................................................... 604

Books Recently Published ................................................................. 616
Thesaurus and Ontology Construction for Contra Dance: Knowledge Organization of a North American Folk Dance Domain

L. P. Coladangelo

Kent State University, College of Communication and Information, School of Information, 1125 Risman Drive, Kent, OH 44243, United States of America, <lcoladan@kent.edu>

Abstract: This case study aims to preserve and disseminate cultural heritage information about the North American community folk dance tradition of contra dance through development of a thesaurus of choreographic terms and a domain ontology. A survey of dance resources was conducted, reviewing historic and modern examples of contra dance choreography notation and instructions, records of dance events, and recordings of dance performances. Domain and content analysis were performed on the resources to collect and organize concepts and themes regarding choreographic components and their relationships, the structure and function of cultural works, their creative expressions, and the evidence of those expressions in documents and recordings. Vocabulary used in the description of contra dance choreography was identified, classified, and notated to build a thesaurus, which was used as the basis of a domain ontology. Ontology building methodology and existing conceptual models for cultural heritage domains guided the ontology development and revision phases. The study also seeks to safeguard an intangible cultural heritage by applying knowledge organization and semantic approaches to folk dance in order to model such challenges as multiple, simultaneous modes of communication and forms of representation, modular conceptual components, descriptive sequences, differing levels of structured information, and complex cultural networks found at various levels of domain discourse.

Received: 11 April 2020; Revised: 5 August 2020; Accepted: 7 August 2020

Keywords: contra dance, domain, ontology

1.0 Introduction and background

Country dancing is an intangible cultural heritage (ICH) of traditional social dance that spans European and North American history from (at least) the seventeenth century to the present day. Contra dance represents a currently active form of country dance performed mostly in the United States and Canada but also worldwide. Information about contra dance occurs in unstructured or semi-structured formats, but there are no standardized vocabularies or domain models in support of a knowledge base. This study, CONTRA (Contra dance ONtological and Thesaural Representation and Application), applied knowledge organization (KO) approaches to safeguard contra dance as an ICH domain. It proposed: 1) creation of a Contra Dance Thesaurus; and, 2) construction of the Contra Ontology, adapting the IFLA Library Reference Model (LRM) as a general framework, along with aspects of the object-oriented Functional Requirements for Bibliographic Records (FRBRoo) model, the DOing REusable MUSic (DOREMUS) data model, and Linked Irish Traditional Music (LITMUS) Ontology.

This paper reports major steps and research findings of the study. Section 1 provides background for ICH safeguarding and the history of contra dance. Section 2 reviews previous research in using KO tools to structure ICH infor-
Contra dance is part of a larger tradition of country dancing, a broad term for a style of community social dancing formed and popularized in England. When country dancing became fashionable in continental Europe in the late seventeenth century, the tradition of dances composed in long, or longways, lines, now influenced by the French, migrated back to England as contradance, or, later, contra dance. These dances were then brought to North America by English colonists, where they took hold in the New England and Appalachian regions. Although its popularity ebbed and flowed throughout the nineteenth century, contra dance experienced important revivals in the early to mid-twentieth century. Today, contra dance and its heritage are largely supported by the Country Dance and Song Society (CDSS), a national organization and a network of affiliated local organizations committed to preserving and continuing the traditions of American folk dance and music.

1.3 Contra dance structure

Country dances are characterized by groupings of dancers executing movements or “figures” in geometric patterns within the dance space. These movements are predetermined by existing choreography describing the included figures, who performs what figures, and how they are to be performed. Contra dancing traditionally consists of long lines of dancers organized into groups of four people/two couples (Figure 1). A caller teaches and prompts the choreography of a given dance, which follows a sixty-four-beat AABB structure (divided into eight-count phrases) matching accompanying folk tunes (Figure 2). As dancers perform the dance each time through, they move up or down the long line to “progress” to a new couple with whom they will interact. This series of figures and progressions keeps the dancers moving as they repeat the same patterns within the AABB structure.

1.4 Research aims and purpose

This study sought to examine KO as a means of preserving the history and traditions of an ICH domain, especially as KOSs and domain models are regarded as ICH safeguarding practices. It explored contra dance as a case study, using its rich vocabulary as the basis of a thesaurus of choreographic terms, which in turn served as the foundation for a domain ontology. It examined challenges and used KO to model an ICH characterized by multiple, simultaneous modes of communication and forms of representation for the same concepts as well as domain knowledge composed of modular conceptual components and descriptive sequences. The use of semantic enhancement strategies powered by ontological and linked data approaches was also investigated to address the challenge of organizing differing levels of structured information and complex cultural networks at various levels of domain discourse.

2.0 Literature review

2.1 KO as a method for ICH safeguarding

The development of ontologies and semantic technologies has supported tools and services that transmit heritage and encourage knowledge building as a safeguarding strategy. For systems to teach Tsamiko and salsa dances, Chantas et al. (2018, 3) cited the efficacy of ontologies for ICH safe-
guarding as “platforms that ... manage to aggregate content of various CH domains in a large-scale database and provide convenient access to that data.” Ontologies have also supported interactive educational systems like i-Treasures (Dimitropoulos et al. 2018) for European intangible culture, an augmented mobile application about Korean cultural heritage sites (Kim et al. 2016), a web application for the Manchu costume culture of China (Huang 2018), and a knowledge base repository for Indian classical dance (Mallik et al. 2011). Another important example, CultureSampo, used semantic technologies to aggregate cultural heritage contents from heterogeneous data sources into a semantic portal. Hyvönen et al. (2009) proposed such a portal for cultural heritage built on a cross-domain infrastructure of ontologies, metadata standards, and related services, resulting in a multilingual end-user application that facilitated exploration and visualization of Finnish heritage.

2.2 ICH domain KOS development

Systems and services for the transmission of ICH need to be supported by underlying KOSs. Classification schemes were proposed by Hu et al. (2014) to organize Taiwanese Indigenous folk dances and by Karavarsamis et al. (2016) for
salsa dance steps. Huang and Huang (2013) used taxonomies in the classification of Taiwanese Indigenous cultural heritage. Furthermore, KO structures (Chansanam and Tuamsuk 2015; Kaewboonma and Tuamsuk 2016), taxonomies (Tuamsuk et al. 2016), and thesauri (Chansanam et al. 2015) have been developed as precursors or intermediate stages toward the construction of ICH domain ontologies. Existing cultural heritage thesauri—like the Art and Architecture Thesaurus (AAT)—have been cited by Alakus et al. (2018) for their relevance in indexing cultural heritage objects with intangible aspects. Baca and Gill (2015) detailed the process of publishing AAT, along with other Getty vocabularies, as linked open data; AAT LOD was later employed by Wijesundara et al. (2016) in the enrichment of Sri Lankan cultural heritage information.

Furthermore, in elucidating potential requirements for systems to preserve dance information, like a dance ontology, Clareance (2015) enumerated aspects such as choreography, performance venue, related tangible objects, as well as meaning and representations in dance forms. Ceusters and Smith (2011) explicitly mention representing the domain in a machine-processable way and methods to support semantic analysis of domain content, including the successive moves in a contra dance.

Just as documentation and digitization alone do not address the safeguarding paradigm, the context-rich environments of cultural heritage domains may not be fully conceptualized by KOSs lacking well-defined semantic relationships. In her discussion on the role of thesauri in the age of the semantic web, Dextre Clarke (2016, 142) posed whether thesauri will evolve toward ontologies due to the “potential for establishing and exploiting differentiated relationships between concepts.” Hjørland (2016, 152) furthered the discussion by stressing the importance of semantic relations in KOS tools, in that different kinds of relations have different importance in different domains.” The need for highly structured semantic descriptions for art objects led Wielinga et al. (2001) to convert a relevant thesaurus, AAT, into an ontology. Moreover, Hyvönen (2009, 762) reminded that a traditional thesaurus may not be “enough from a semantic viewpoint” since “its semantic relations have been constructed mainly to help the indexer in finding indexing terms, and understanding the relations needs implicit human knowledge.”

Through more detailed structuring of semantic descriptions, ontologies present a bridge between human conceptualization of a domain and its representation as machine-processable information in the semantic web. Eide and Ore (2018, 182) defined an ontology in the context of cultural heritage as “a special kind of data model dealing with formalized conceptualizations.” Doty (2013, 1) explained that an ontology can “identify what is essential about a domain of knowledge and to distinguish among those essential elements ... to represent such knowledge in physical form.” This process of formalizing and representing ICH knowledge makes it tangible (insofar as it can be explored, shared, reused, and analyzed) outside of the minds and practices of practitioners and knowledge-bearers, thus supporting its safeguarding.

2.3 Conceptual models for cultural heritage and their use in ICH domain ontologies

2.3.1 CIDOC CRM

CIDOC CRM is a “formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information” (Doerr 2009, 468), extensively used in modeling various ICH domains. It was the basis of ontologies developed by Dou et al. (2018) for the 24 Solar Terms of China, Goienetxea et al. (2012) for folk song metadata, Hu et al. (2014) for the Pang Wang Festival of the Yao people of China, Martini et al. (2016) for the personal narrative ontology OntoMP (Ontology for The Museum of the Person), and Tan et al. (2009) for the Funeral Dance of the Tujia people of China. As an event-based conceptual model, CIDOC CRM also influenced the development of local ontologies like the drama ontology Drammar (Lombardo et al. 2016). The Linked Art Data Model (2020) used a “streamlined” version of CIDOC CRM with the Getty Vocabularies as value vocabularies. Despite its ubiquity, some studies have criticized the model for being too “museum-centric” (Brownlow et al. 2015, 5), citing the limited use and relevance of its expansive number of entities and a lack of concepts that would be important in modeling particular ICH domains (Pramartha and Davis 2016).

2.3.2 FRBR

FRBR is a conceptual entity-relationship model “that identifies and clearly defines the entities of interest to users of bibliographic records, the attributes of each entity, and the types of relationships that operate between entities” (IFLA Study Group on the Functional Requirements for Bibliographic Records 2009, 3). Doerr (2009, 471) described its “innovation ... to cluster publications and other items around the notion of a common conceptual origin—the ‘Work’, in order to support information retrieval.” This conceptual structure of group one entities (Work, Expression, Manifestation, Item, or collectively, WEMI), represent a hierarchy of “products of intellectual or artistic endeavour” (IFLA Study Group on the Functional Requirements for Bibliographic Records 2009, 13). In terms of ICH domain research, it has been applied extensively to the performing arts, like the Music Ontology (Raimond et al. 2007), and a metadata standard for Greek folk dance (Giannoulakis et al. 2018).
2.3.3 **FRBRoo**

In an effort to improve the organization of cultural heritage information between museum and library communities, a harmonization project between *CIDOC CRM* and *FRBR* led to the creation of *FRBRoo*, a “formal ontology intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information” (Working Group on FRBR/CRM Dialogue 2016, 12). Doerr (2009, 472) described it as “a realistic, explicit model of the intellectual creation process” well-suited for the performing arts, as it differentiates and connects creative works, interpretations of those works, and their recordings with their symbolic forms and physical carriers. *FRBRoo* was utilized by Marolt et al. (2009) for the Ethno-Muse project of Slovenian folk music and dance, El Raheb and Ioannidis (2014) for an ontology to model digital libraries of dance information, Monika et al. (2017) for the Kecak dance of Indonesia, and Park et al. (2019) for modeling traditional performing arts archives in South Korea. Le Boeuf (2012) also suggested its use in linked data applications for performing arts; two such implemented examples included the work of Lisena et al. (2018) in the DOREMUS project for the enrichment and reuse of music datasets, and the LITMUS ontology by Weissenberger (2017) for traditional Irish music.

2.3.4 **LRM**

In 2017, IFLA consolidated and updated, and thus deprecated, the FR family of models through the creation of *LRM* (Riva et al. 2017). *LRM* enhanced the conceptual entity-relationship model in *FRBR* through its WEMI framework, indicating degrees of flexibility in defining entity requirements and attributes. Although there has not yet been a direct alignment of *LRM* in modeling an ICH domain (its intended community is library-centric), the creators of the DOREMUS model noted that their conceptions of Work and Expression (in their extension of *FRBRoo*) were actually closer in alignment to those same entities as presented in *LRM* (Lisena et al. 2018).

3.0 **Materials and methods**

3.1 **Contra dance historical and choreographic resources**

No standardized resources or definitive compendiums of choreographic instructions exist for contra dance. Information regarding contra dance terminology, creative works, and cultural history is diffuse and unstructured, existing in many forms and formats, such as printed texts, online sources, audio and video recordings, dance cards collected by dance callers, and even in the minds and memories of domain practitioners. The difficulty of gathering information about contra dance choreography is further compounded by the very nature of it being an actively practiced, intangible heritage. Organizing the domain is challenging because of differences in dance vocabulary and inconsistencies in traditional terms used among communities and regions (Murphy and Murphy 2019). Some historical or traditional dances, called “chestnuts” (Millstone 2002), may be well-documented or reprinted in many places, but those instructions or notations have not always been written or recorded in the same way, or may have been altered over time by the folk process. Moreover, new dance works and new elements of choreography are being written even into the present. The practice of adding new examples to an existing repertoire makes a definitive accounting of the totality of contra dance choreography an impossible task.

Notwithstanding the complexity present in the addition and evolution of dance works, there are significant challenges related to vocabulary control with regard to domain knowledge, especially when written choreography for the same dance can appear markedly different depending on the nature or audience of a given resource. A choreographer may write detailed instructions that would be valuable to teachers and dance callers. These same instructions might then be shortened or notated to be collected and anthologized. Choreography may be even further abbreviated to fit on a caller’s index card or notated to match their calling style. As a result, contra dance choreographic information and its accompanying vocabulary, even for the same dance work, can appear notated in a variety of ways and at varying levels of detail throughout the domain discourse.

With those relative complications in mind, this study looked to apply KO methods and systems to mitigate the difficulties inherent in safeguarding the ICH due to the heterogeneous nature of contra dance information. In the development of a contra dance thesaurus and subsequent ontology, the study took its literary warrant (Barité 2018) for establishing choreographic terms and domain entities through a variety of available resources. These resources included: 1) the digitized syllabi and the index compiled by Smukler (2014) for the Ralph Page Dance Legacy Weekend Collection, 1988-2017; 2) printed and edited anthologies and histories by Dart (1995), Gunzenhauser (1996), Smukler and Millstone (2008), and Pittman et al. (2009); 3) choreography published in *CDSS News* between 2010 and 2019; 4) online databases, glossaries, and indexes, like those compiled by Owen (2003) and Gascon (2015), as well as the crowdsourced ContraDB (https://contradb.com) and The Caller’s Box (http://www.ibiblio.org/contradance/theccallersbox), compiled by Chris Page and Michael Dyck; 5) various monographs and manuals by well-known choreogra-
phers and callers; 6) electronic mailing lists and blog posts; 7) video recordings of performances and events on social media websites like YouTube; 8) research articles studying contra dance as a historic, folkloric practice; and, 9) research articles on the application of other disciplines, such as mathematics, to the study of country dance.

3.2 Methods of analysis

3.2.1 Domain and content analysis

Domain and content analysis of contra dance involved an examination of domain discourse within historical documents (texts, recordings, etc.) with an iterative process of analysis rooted in grounded theory (Thornberg and Charmaz 2014), to identify, contextualize, and refine conceptualization of the vocabulary, works, expressions, identities, and entities that were important parts of the dance tradition. Domain analysis was guided by Hjørland and Albrechtsen (1995) and refined further through Tennis (2003) by identifying its areas of modulation, or extension and name (the history, tradition, practices, and evidence of contra dance as an ICH), and its degrees of specialization, or focus and intersection (contra dance as a primarily North American form of country dance but distinct from other forms like English country dance, traditional square dance, and modern western square dance). Domain analysis toward ontology building for an ICH domain was also considered to meet the following criteria of Hjørland’s (2002) approaches by: 1) providing access to information sources as a gateway to an ICH subject; 2) constructing a special classification system to organize an ICH domain in order to understand semantic concepts; 3) supporting information retrieval; 4) contributing to the historical study of ICH through organization of traditions and forms of expression; 5) developing information architecture that represents the domain’s inherent organization; 6) organizing ICH knowledge in a paradigmatic way; 7) studying discourse, language, and semantics within an ICH community; and, 8) revealing mental or cognitive models of an ICH domain.

In employing content analysis, this study also looked toward Mayring (2000) in conducting a qualitative analysis that focused on the manifest content of recorded material within the domain, as well as its themes, main ideas, contextual information, and formal aspects. Communicative content was broken down into component parts and categorized for further analysis, identifying the occurrence and frequencies of terms (word counts) to situate concepts in context (Drisko and Maschi 2016). Ultimately, three broad areas were found within the domain: 1) choreography (vocabulary, components, and sequences); 2) creative works and their evidence (dances, notations, documents, and performances); and, 3) cultural networks and history (people, groups, places, timelines, and events). Notably, similar methods for identifying key concepts within ICH domain resources for KOS development have also been employed by Lombardi et al. (2018) in representing drama and dramatic narrative and by Kaewboonma and Tuamsuk (2018) and Tuamsuk et al. (2016) to organize folk traditions of the Greater Mekong Subregion.

3.2.2 Faceted analysis toward thesaurus development

Within the process of domain analysis, broad yet mutually exclusive categories, or facets, may also emerge. The process of faceted analysis and classification used in this study generally followed those delineated by Vickery (1966), Mills (2004), and La Barre (2010), as elucidated by Hjørland (2013). Faceted analysis was previously utilized in the KO of ICH by Madalli et al. (2015) for a music domain and Sca-turro (2013) for performing arts. Additionally, Dai et al. (2014) applied a theory of knowledge classification to ICH in order to express semantic relationships using a faceted structure more popularly recognized as the five Ws: 1) when; 2) where; 3) what; 4) why; and, 5) how. Construction of a faceted thesaurus for choreographic terms in this study was also guided by design approaches and insights from Aitchison et al. (2002), Broughton (2006), and Tudhope and Binding (2008).

3.3 Ontology building methodology

Although ontology building methodologies vary (Gómez-Pérez et al. 2004, Stuart 2016), ontologies and linked data projects for ICH domains developed by Kaewboonma and Tuamsuk (2018), Pattuelli et al. (2015), Tuamsuk et al. (2018), and Weissenberger (2017) have employed strategies based on those defined by Noy and McGuinness (2001). In their primer for ontology development using Protégé software (Musen, 2015), which was used in this study, Noy and McGuinness (2001, 4) proposed a “simple knowledge-engineering methodology,” acknowledging that ontology construction is an “iterative process” and may include “viable alternatives.” Using stages similar to those from Noy and McGuinness and previous ICH ontology studies, the following steps were taken: 1) determination of the scope and domain; 2) selection of data sources; 3) analysis to establish potential vocabularies and authority files, as well as initial constraints of a domain model; 4) identification of key concepts and relationships through the use of KOS like taxonomies, thesauri, and semantic networks to aid model building; 5) examining other cultural heritage ontologies for reuse, extension, or adaptation; 6) formally naming and defining classes, their properties, as well as domains, ranges, and data types of properties; 7) population of the ontology with collected sample data to create instances; 8) reevaluation
and revision of the ontology prompted by the integration of instances; and, 9) testing using simple SPARQL (SPARQL Protocol and RDF Query Language) queries to determine basic anticipated functionality.

4.0 Results

4.1 Contra Dance Thesaurus construction

The first stage of the study resulted in a thesaurus for contra dance choreographic and cultural terms. In describing her process of teaching dances efficiently within the limitations for a walkthrough (the introduction to a dance’s instructions before the music and calling begins), nationally known dance caller Beth Molaro explained (Merritt 2014) the following structure: “who you’re doing something with, what you’re going to do with them, and where you end up.” Notably, this same type of faceted classification was used by Scaturro (2013), citing Ackoff (1989), through the fundamental questions (what? when? where? who?) to be answered in transforming data into information; thus, forming facets as answers to these “W” questions closely followed Molaro’s own evaluation of the faceted essence of contra dance choreography. Consequently, the vocabulary of dance instructions fell into one of these structural areas: roles (who), figures (what), durations (when or how), and directions or distances (where or how). From these facets, superclasses were formed (Figure 3) and terms were classified to create a controlled vocabulary for concepts belonging to: 1) particular figures or movements; 2) the roles performed by individual dancers and groups of dancers; 3) the directions indicating where dancers move or how to execute a figure; and, 4) the standard durations, fractional portions, number of places, or lengths of (musical) time a dancer will perform a figure. Further hierarchical classifications were made based on similarities in structure and function, as well as relationships to other facets. For instance, figures were ordered by the maximum number of dancers involved in the execution of a particular movement (a figure’s relationship to “who”), and then by the nature of their movement (“where” or “how”), such as directionality. In contrast, durational and distance terms were classified by the convention used in counting or marking that information (e.g., fractions of a full figure, number of beats of music, or number of places moved on the dance floor). Additionally, other classes were necessary for organizing vocabulary that described structural elements like choreographic formations and methods of progression, as well as cultural terminology. The thesaurus was then finalized by delineating associative relationships between related terms, defining scope notes for preferred terms (Figure 4), noting terms, and implementing both an alphabetical index and a classified schedule.

4.1.1 Challenges and limitations of thesaural representation

Although the thesaurus defined relationships between terms, these relationships lacked the semantic complexity and specificity necessary for a full conceptualization of the domain. For example, it was found that single axis classification was a challenge in organizing terms for figures that captured comparative similarities and differences in their attributes (Figure 5). Because contra dance choreography is based heavily on the relative positions of dancers and their direction of momentum, the domain model required definition of these attributes to properly structure the relationships between figures, the roles involved, and valid applicable terms for modifying direction, duration, or distance. Choreographers and callers, in particular, need to un-
stand these attributes to make creative decisions about how dance movements fit together and how to build satisfying dance programs; therefore, a vocabulary classified along any single property axis or using a sole characteristic would be incomplete. Additionally, choreographic dance works themselves possessed complex links to other entities such as formations, levels of difficulty, sequences of calls, matching or suggested tunes, agents like choreographers, and other contextual domain information that could only be represented as properties of these works through ontological modeling.

There were also challenges that arose in the presence of multiple preferred terms and the absence of preferred terms. The use of gendered language in the contra dance community for the dance roles traditionally performed by men and women has been increasingly disputed, especially by dancers who identify as LGBTQ and those who believe the explicit gendering of roles is irrelevant to the performance of the
dance or exclusionary to those who wish to dance both roles. This has led to the adoption of different gender-neutral terms for the same role concepts (Figure 6). The use of these terms, however, can differ between local communities, and in fact, within the same community, as some communities support the use of multiple pairs of terms by allowing whatever role convention a particular caller prefers or by designating some dance events as gender neutral. Conversely, use of the term "gypsy," which is used as the traditional descriptor for a particular figure, has been deeply criticized of late for its racist history as an ethnic slur. Some callers, organizers, and communities have prohibited the term while others have defended it. At present, many alternatives have been proposed (Figure 7), but no consensus has been reached. It has been left to individual callers and local communities to make ad hoc decisions whether to retain the traditional term or to adopt a different one (German et al. 2019). As a result of these gender- and race-based critiques of traditional terminology, it was found that a KOS for contra dance needed to support the simultaneous use of multiple terms for the same concept and to contextualize the history of these terms within the tradition.
4.2 Contra Ontology construction

4.2.1 Classes and properties

The second stage of the study resulted in the development of a domain ontology. In addition to the choreographic vocabulary, entities like dance works, notated or written dance instructions, people and organizations, performances and events, and other concepts endemic to the domain were gathered for inclusion and placement into the ontology. Addition to and refinement of classes from the thesaurus also took place at this stage (Figure 8). The process then moved to the assignment of properties for class entities and their attributes, along with delineating valid class domains for properties and class ranges for acceptable values of properties to support ontological inference. The resulting ontology described attributes/properties of choreographic elements, components, and sequences, dance works, versions of notation of dances, called versions of dances, dance performances, evidence/documentation of performances, cultural events, and cultural practitioners and organizations. It also described relationships between entities by linking and connecting creative works to other works, works to their various forms of expression, expressions to their various forms of documentation, works, expressions, and documentation to people, places, and events, people to other people, people to groups, groups to other groups, people and groups to events, and events to other events.

One of the challenging aspects to model was the nature and structure of the choreographic instructions, both conceptually and in practice. Although information for figures, roles, directions, and distances/durations were organized separately in the thesaurus, in actuality, these vocabulary components are (re-)combined for the purposes of constructing a dance call, or the complete string of instructions for executing a portion of the dance. A figure is a necessary component of a call, but it is also modified or clarified by terms from the other component classes. Keeping in mind Molaro’s faceted calling instructions, the basic structure for a call is a combination of a figure + dance role + direction + distance/duration, depending on the nature of the call and the valid types of components that can be used as values for the properties of a given figure. A class of entities Call was developed to house the various combinations of choreographic concepts that compose specific strings of instructions by assigning those components as values of properties of Call (Figure 9).

Another challenge was representing the order and duration of calls in a dance work, as the execution of a piece of choreography must correspond with the accompanying music (sixty-four-beat tunes in AABB structure). To capture the basic AABB sequences, properties a1, a2, b1, and b2 were created for Dance entities. If more detailed sequence information was required, two possibilities were posited for different circumstances: 1) if the exact duration of the calls was unknown or estimated, instructions would be subdivided as a1.1, a1.2, a1.3, etc., to show the order of calls without regard to the number of beats; and, 2) sections could be further subdivided as a1.1-2, a1.1-4, a1.1-6, up to
4.2.2 Reuse of existing conceptual models and alignment to other domain ontologies

As the bottom-up analysis identified entities and their relationships through thesaurus construction and the initial stages of ontology development, a top-down approach concurrently looked at established models that would clarify and refine the domain ontology with an eye toward future interoperability. Three considerations emerged when evaluating the applicability of existing models to the domain: 1) ease of use through structural equivalence; 2) semantic interoperability through very close or exact matching of domain entities to the definitions of their counterparts in an existing model; and, 3) use of current models and standards, including novel approaches. From these perspectives, CIDOC CRM was indeed found to be too unwieldy and expansive to be wholly applicable to contra dance, yet two existing conceptual models, FRBRoo and LRM, were well-suited to the overall structuring of the ontology. Two additional domain ontologies based on FRBRoo, DOREMUS and LITMUS, aided in the development of specific domain classes (Figure 10).

4.2.2.1 Structural equivalence with FRBRoo and LRM classes

FRBRoo contributed domain classes for works, expressions, and performances, and LRM reinforced integration of the WEMI structure. The classes (and subclasses) of CRM re-configured in FRBRoo to define and classify works, expressions, and performances were deemed appropriately equivalent. Work and Expression in LRM matched domain classes for conceptual works of choreography and their various forms of expression (whether written/notated, spoken, or physically performed), respectively. Manifestation was suitable to classify documentation or recordings of domain expressions. LRM also provided a generic Agent, with subclasses for individual people (Person) and groups or organizations (Collective Agent), both of which exist in contra dance.

4.2.2.2 Semantic interoperability with LRM and with FRBRoo and its domain extensions DOREMUS and LITMUS

The contra dance domain also included entities semantically aligned to those in LRM and FRBRoo. In LRM, the Work entity is the “intellectual or artistic content of a distinct creation” (Riva et al. 2017, 21) which comes into existence by virtue of a first Expression. Dances and Tunes, as subclasses of Works, come into being through the notated instructions of NotatedDances or as Performances (expressed by human movement or by sound). These forms dovetailed with LRM-E3 Expression, which is a “distinct combination of signs conveying intellectual or artistic content” (23), which includes “visual, aural or gestural signs.” Expressions are then captured onto a carrier known as a Manifestation (LRM-E4), classified in the domain by either AnalogManifestation or DigitalManifestation. Although the contra dance model did not need to extend down to the Item (LRM-E5) level, in some cases a specific Manifestation (e.g., a caller’s dance card) could be both an Item, as well as a one-of-a-kind Manifestation (again, the carrier of the dance card itself) of an Expression (a NotatedDance) of a Work (Dance), thus adhering to the same semantic structure for those entities as LRM.

FRBRoo was also semantically interoperable, beginning with F1 Work, comprising artistic and intellectual concepts/ideas, emphasizing the role of people or collective agents in the execution or elaboration of a Work, especially through Expressions (Working Group on FRBR/CRM Dialogue 2016). The F20 Performance Work was also applicable as a set of “concepts for rendering a particular or a series of events” (2016).
of like performances” (67), which encapsulated the abstract content of a Dance used to guide future Performances. Much as in LRM, F2 Expression comprised “the intellectual or artistic realisations of works ... such as texts ... musical or choreographic notations, movement pattern, sound pattern...or any combination of such forms” (55). This definition accorded with NotatedDances as Expressions, which give choreographic instructions for Dance concepts. A NotatedDance also qualified as an F25 Performance Plan, which comprised “sets of directions to which individual

<table>
<thead>
<tr>
<th>Contra Ontology classes</th>
<th>LRM classified entities</th>
<th>Classified entities from FRBRoo (F) / CRM (E) and FRBRoo extensions through DOREMUS (M) and LITMUS (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work</td>
<td>LRM-E2 Work</td>
<td>F1 Work</td>
</tr>
<tr>
<td>Dance</td>
<td></td>
<td>F14 Individual Work or F20 Performance Work M44 Performed Work T14 Dance</td>
</tr>
<tr>
<td>Tune</td>
<td></td>
<td>F14 Individual Work or F20 Performance Work M44 Performed Work T10 Instrumental Tune</td>
</tr>
<tr>
<td>Expression</td>
<td>LRM-E3 Expression</td>
<td>F2 Expression or F22 Self Contained Expression</td>
</tr>
<tr>
<td>NotatedDance</td>
<td></td>
<td>F25 Performance Plan</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td>F31 Performance T16 Dance Performance T36 Music Performance M42 Performed Expression Creation M43 Performed Expression</td>
</tr>
<tr>
<td>TuneSet</td>
<td></td>
<td>T11 Instrumental Tune Set</td>
</tr>
<tr>
<td>Call</td>
<td></td>
<td>F23 Expression Fragment T17 Dance Component</td>
</tr>
<tr>
<td>Manifestation</td>
<td>LRM-E4 Manifestation</td>
<td>F3 Manifestation Product Type or F4 Manifestation Singleton</td>
</tr>
<tr>
<td>Agent</td>
<td>LRM-E6 Agent</td>
<td>E39 Actor</td>
</tr>
<tr>
<td>Person</td>
<td>LRM-E7 Person</td>
<td>F10 Person/E21 Person</td>
</tr>
<tr>
<td>CollectiveAgent</td>
<td>LRM-E8 Collective Agent</td>
<td>F11 Corporate Body</td>
</tr>
<tr>
<td>Band</td>
<td></td>
<td>T32 Band</td>
</tr>
<tr>
<td>PersonnelRole</td>
<td></td>
<td>T27 Role (equivalent to / subclass of F10 Person)</td>
</tr>
<tr>
<td>Event</td>
<td></td>
<td>F8 Event</td>
</tr>
<tr>
<td>Place</td>
<td>LRM-E10 Place</td>
<td>F9 Place / E53 Place</td>
</tr>
<tr>
<td>Time</td>
<td>LRM-E11 Time-span</td>
<td>E52 Time-Span</td>
</tr>
</tbody>
</table>

Figure 10. Contra Ontology classes mapped to their semantic and structural equivalent classes from LRM, FRBRoo, DOREMUS, and LITMUS (Coladangelo 2020, 94).
performances ... should conform” (71). Furthermore, an F31 Performance involved “activities that follow the directions of a performance plan” (75), included in the domain model through a property of Expression that related one Expression to another (relatedExpression), allowing the NotatedDance to be connected to a related Performance.

A Performance in the domain model, then, was an Expression of a Work situated at the nexus of Works, Expressions, Events, and Agents. In contra dance, a Performance is not a demonstration for an audience, but the realization of a dance work within a communal setting in which conceptual dance instructions are expressed as physical signs by the dancers’ movements, a caller’s words, and accompanied by live music. The physical movements mark the Performance, but the translation of dance concepts into movement is not the sole component (even if it is the defining element), because the realization of contra dance performances includes the efforts of callers, musicians, and others taking place in a certain time and space. Alignment to other models meant that Performance needed to be linked to (or classified as) events, which accorded with F8 Event, where Events and Performances occur in Places (F9 Place, CRM E53 Place, and LRM-E10 Place). This usage was confirmed by applying the same logic of the M43 Performed Expression class in the DOREMUS model, which reinforced the Work-Expression-Event relationship and explicitly advanced the notion that a performance could give rise to or constitute an expression and an event, sharing properties of each of those classes.

The DOREMUS and LITMUS models furthered understanding of the relationship between expressions, performances, and related entities. This was structured through F28 Expression Creation in its intermediary role creating an F22 Self-Contained Expression and as a realization of an F14 Individual Work, consisting of an E7 Activity carried out by a E21 Person with an M31 Actor’s Function and/or M32 Actor’s Responsibility. This accorded with a Person (or Agent) that inhabits or performs a particular PersonRole in the domain ontology, for which LITMUS provided analogs (e.g., T27_Role class, with instances like I11_Dance_Callner, I23_Composer, and I30_Musician). LITMUS bolstered the domain conceptualization of choreographic elements with class T17_Dance_Component (as in elements or portions of dance choreography) which aligned with instances of the Figure, Role, Direction, and DistanceDuration classes composing a Call. The LITMUS representation of sets of Tunes played together (T11_Instrumental_Tune_Set) was appropriated in the domain ontology as the TuneSet class and was extended to address the challenge of the existence of PerformanceSets by describing groups of dances called within the same session (a DanceSet), or in a single performance of dances called one right after another (a Medley).

4.4.2.3 Novel approach in leveraging LRM Res and Nomen classes

Lastly, the domain ontology proposed ways to leverage the innovations of the LRM Res and Nomen classes to model and trace the lineage of tradition and thematic content. Res is “any entity in the universe of discourse” (Riva et al. 2017, 20), providing a superclass to represent any other entity or concept that will be important to capture but has not been classed or named. One of the attributes available to Res is a Category (LRM-E1-A1), which is “a type to which the res belongs” (40). Res and its Category attribute addressed the challenge of representing lineage, like a “family” of dances, to group entities that share some historical or traditional conceptual source. Using this framework, the ontology was revised to identify a class for linking Works that at one time shared an origin, provenance, or common root. A second challenge surmounted by the Res class would be representation of thematic content or subject matter not structurally related to dances but important to the history of choreography development. This could be represented through the Category attribute or by a more structured version of the Res attribute Note (LRM-E1-A2), which provides information that “is not recorded through the use of specific attributes and/or relationships” (40), such as detailed narrative or descriptive information. This contextual information would be useful to callers and organizers planning dance programs and events. It may also encourage historians and researchers to study trends of cultural interest to the contra community.

Moreover, the Nomen entity (LRM-E9) can be linked to other entities by appellation relationships, meaning that a Nomen can be related as a name or label for an entity but can also possess its own attributes. This flexibility would support multilingual representation in the broadest sense of the term, in that a Nomen can have its own attributes (or properties) for language and script, but can also be represented and described by various levels of abstraction, signs or symbols, verbal utterances, physical gestures or movements, or clips of recorded media. A Nomen could further support tradition and lineage by having associated authority files, as well as attributes for the Place, Event, or Time in which it was actively used. Because contra dance possesses a rich choreographic vocabulary, the Nomen for each instance of a dance component could provide a wealth of contextual information regarding the cultural history of each vocabulary term.

4.2.3 Support for different levels of domain discourse

A significant challenge was the ontological representation of the different kinds of discourse inherent in the concepts and relationships between dance works, the notation of
dance instructions, the verbal cues of dance calling, and the physical performance of dances. As a result, the ontology was revised to include distinct but related types of expressions (NotatedDance, CalledDance, and Performance) of a Dance, which are documented through examples of manifestations. A NotatedDance represents instructions in a written form. Examples of dance notation for the same set of instructions can vary significantly (Figure 11), but the manner of notation represents an important aspect of cultural communication within the domain. Similarly, in a CalledDance, the spoken/verbal instructions from a caller represent another vital form of discourse that shows personal styling and cultural influences in the choice of both instructional language and supplemental phrases or “patter” (Parkes 2012). Furthermore, a Performance is a form of discourse as a physical or bodily expression connecting practitioners, creative works, related expressions, locations, time periods, and events. These forms of expression also leave behind evidence of their cultural practices or manifestations of the cultural heritage that can be examined as part of the domain discourse. To provide a complete understanding of these levels of cultural representation and communication, the ontology needed to account for the conceptual, practical, and evidential levels of heritage as signified in the domain.

### 4.2.4 Semantic annotation of Calls

Another challenge was ontologically representing different levels of specificity in the structuring of Calls. For example, natural language, alternate vocabulary, or implied or omitted instructions were found in CalledDances and NotatedDances, so that information was conceptually identical to its original work but practically different. Because these expressions form an important part of the domain discourse, it would have been inappropriate to control their vocabulary or standardize their language to make them conform to a single structure. Instead, it was found that these less structured expressions could be semantically enhanced by annotating their constituent Calls by assigning values to properties for named/known components of Figures, DanceRoles, Direction, and Distance/Duration. This led to ontological modeling of different subclasses of Calls (UnstructuredCall, SemiStructuredCall, and StructuredCall) depending on the presence of certain levels of semantic specificity (Figure 12). Semantic annotation addressed the challenge of preserving important historical and cultural variations in domain discourse while enhancing disparate or less structured forms of dance information to be connected meaningfully in the ontology through equivalence assertion or ontological inference.

### 4.2.5 Domain knowledge base support and future testing

Once the ontology was populated with sample instances, ontological inference and simple SPARQL queries were used to test its basic potential to support a knowledge base. With properly structured class definitions, domains and ranges for object and data properties, and assigned property values, an ontological reasoner was able to properly classify examples of instances by inference. Examples of these included inferring classifications of figures for two dancers and tune sets as well as an intersection class for dances in duplet improper formation. SPARQL, by contrast, only queries asserted, not inferred, knowledge, meaning that data in RDF triples must be explicit to be returned as results. Examples of SPARQL queries conducted included finding the name of a dance (“Chorus Jig”) that is set to a tune that has an alternate title “The Glen Road to Carrick” and returning a list of choreographers based in Ohio, where the ontology was populated with four asserted instances of choreographers (Don Armstrong, Becky Hill, Carol Kopp, and Tanya Rotenberg) and three instances of Agent based in Ohio.

---

**Figure 11.** Examples of differently notated versions for the same choreographic concepts (Coladangelo 2020, 63-4).
(Becky Hill, Carol Kopp, and Dick Swain). Although additional population with instances and more complex SPARQL querying would enhance its future usability, the ontology showed valid, basic functionality as a nascent domain knowledge base. Further research into functionality and knowledge base support would include evaluation of the ontology with different user groups of domain practitioners like choreographers and callers, as well organizations like CDSS, to encourage revision and refinement of the ontology to meet needs related to information retrieval and historical preservation.

5.0 Conclusion

The CONTRA study applying KO approaches to the community folk dance tradition of contra dance resulted in the construction of a choreographic thesaurus and a domain ontology. The study concluded that vocabulary control, ontological modeling, and semantic technologies were well-suited to structure information about contra dance, and the ontology would serve as the infrastructure for a knowledge base to safeguard and disseminate contra dance history and culture. The safeguarding paradigm would be met through organization, storage, search, and retrieval of domain knowledge, including dance vocabulary, choreographic instructions and notations; performance details, name authorities for titles, people, and corporate bodies, places, and events, contextual information, and cultural concepts. The ontology posited a structured data core to which less structured and semi-structured data that characterizes the domain. In this way, the model addressed the challenge of representing different levels of cultural expression while maintaining structured representation of the domain. Future interoperability was supported by semantically structuring domain information in a linked data environment and through alignment with existing cultural heritage conceptual models and other domain ontologies.

Insights were also gained regarding the unique requirements of ICH domains with avenues for future KO research toward ICH safeguarding. Concepts that emerged from this study identified important representational benchmarks or guidance for KO development for ICH domains with the following aspects: 1) modularity of components from the most basic to the highest aggregated level, including intermediary stages or combinations of components; 2) sequences, timelines, or the order of events or entities; 3) differing levels of conceptualization, instantiation, and domain discourse; 4) simultaneous support or validation of multiple, alternate forms of signs, languages, or notations for similar or identical concepts; and, 5) complex contextual information, relationships, and networks of meaning.

Semantic strategies used in KO construction for the formalization of the unstructured and disparate information of contra dance could be applied to adjacent and related folk dance domains, especially the choreography and cultural networks of other country dance domains. This same work could be extended to models related to North American folk music traditions like old-time music or traditional Quebecois music. It could also model cultural domains marked by prescribed, performative, and ritual movement, such as narrative choreography, martial arts, exercise routines, or religious ceremonies. KO of country dance and sequential movement-based ICH would support future implementation of knowledge bases, metadata schemas, semantic analysis tools, linked data approaches, and mapping to other notation methods and domain models. KOSs for these do-

Figure 12. Levels of semantic structuring of Calls (Coladangelo 2020, 102).
mains would also further safeguarding practices like educational training and evaluation systems, remote interaction with cultural performances, augmented and virtual reality applications, and enhancement of user experiences and knowledge building activities for ICH.

References


Domain Analysis Applied to Online Graffiti Art Image Galleries to Reveal Knowledge Organization Structures Used Within an Outsider Art Community

Ann M. Graf

Simmons University, 300 The Fenway, Boston, MA 02115, <ann.graf@simmons.edu>

Abstract: Domain analysis is useful for examination of individual spheres of intellectual activity, both academic and otherwise, and has been used in the knowledge organization (KO) literature to explore specific communities and uses, including web pornography (Beaudoin and Ménard 2015), virtual online worlds (Sköld, Olle 2015), gourmet cooking (Hartel 2010), healthy eating (McTavish 2015), art studies (Ørom 2003), the Knowledge Organization journal (Guimarães et al. 2013), and domain analysis itself (Smiraglia 2015). The results of domain analyses are useful for the development of controlled vocabularies, taxonomies, ontologies, metadata schemas, and other systems for the documentation, description, and discovery of resources, as well as for knowledge discovery in general (Smiraglia 2015; Hjørland 2017). This research describes a methodology for the elucidation of knowledge organization systems (KOS) currently in use on image websites that document graffiti, graffiti art, and street art around the world.

1.0 Introduction

Traditional art museums have historically documented physical collections of artworks held within individual institutions. Museums now often digitize parts of their collections for various reasons: for internal documentation to track condition and maintenance, inclusion records for curated exhibits, and insurance purposes, but also more recently to provide electronic access to images of works online for those who wish to view them without traveling to the physical museum or gallery. Graffiti art, in opposition to traditional museum or gallery works, is created on or placed on the streets themselves, in a gallery of sorts largely by and for the public (Austin 2010; Riggle 2010). The works are, therefore, situated outside the monetized structure of the formal museum or gallery and defy, in their situation and means of creation, the power, ownership, and curation ideals imposed by these institutions. As such, they also largely avoid the careful and controlled documentary practices by such institutions (Graf 2018).

This does not mean graffiti works are not documented, organized, and described. Far from it! Hundreds of websites exist that are dedicated to sharing image galleries of graffiti, graffiti art, and street art from around the world. Hundreds, if not thousands, of image galleries also exist on social media platforms such as Instagram, Flickr, and Facebook, where a range of people, from the artists themselves, to professional photographers, casual observers, and even disgruntled property owners, share images of the works online. What is lacking in these extra-institutional collections is consistency in the descriptive depth and the terminology used to describe the processes involved to produce the works and images of the resultant works themselves (Graf 2018). There is even disagreement on what to call the works, as seen by the overlapping and often contested terms graffiti, graffiti art, and street art (Austin 2010).
The navigational structure and page labels from 241 graffiti art websites were examined to determine the aspects used to organize graffiti art images. Use of various categories of description was tracked across all of these examined aspects of all websites, revealing which were most commonly used around the world. A motivational dichotomy was also apparent, revealing a significant difference in terminology used in “about text” between websites that were determined to be internally or externally motivated. Externally motivated sites are those focused on documentation of graffiti artworks by various artists, while internally motivated sites are those curated by artists themselves with a stated purpose of sharing works to gain commissions, sales, gallery representation, and other forms of direct economic benefit.

The research provides for terminological enrichment of controlled vocabulary tools for use with graffiti art, such as the Getty Research Institute’s Art and Architecture Thesaurus (AAT). It also provides a base for the creation of a metadata schema or application profile of a schema already in existence to better document graffiti art images from four broad groups of aspects for description: general, supports, styles, and locations. These groups for description are the starting point for building a taxonomy of terms relating to graffiti art, with special attention paid to the idiosyncrasies of producing artworks illegally and on various (often) publicly seen but privately owned property. This work adds to vocabularies already in use for the documentation of more traditional art forms and gives voice to the graffiti art community and how this community speaks of graffiti art processes and products. The research is also valuable as a methodological example of domain analysis using evidence from the organizational structures of online image galleries.

1.1 Background

Research in knowledge organization (KO) has shown the value of domain analysis to reveal epistemological evidence of community practice, especially concerning language use within and for individual domains. It is useful for examination of individual spheres of intellectual activity, both academic and otherwise, and has been used in the KO literature to explore specific communities and uses, including web pornography (Beaudoin and Ménard 2015), virtual online worlds (Sköld 2015), gourmet cooking (Hartel 2010), children’s information seeking (Beak 2014), healthy eating (McTavish 2015), art studies (Ørom 2003), the Knowledge Organization journal (Guimarães et al. 2013), and domain analysis itself (Smiraglia 2015). The results of domain analyses are useful for the development of controlled vocabularies, taxonomies, ontologies, metadata schemas, and other systems for the documentation, description, and discovery of resources, as well as for knowledge discovery in general (Smiraglia 2015; Hjørland 2017). This research describes a methodology for the development and enhancement of knowledge organization systems (KOS) used to document graffiti, graffiti art, and street art by way of photographic images of the works.

The academic literature on graffiti is vast and comes from a variety of viewpoints, including that of art, art history, sociology, criminology, urban studies, and history, to name a few. While none, save Gottlieb (2008), discussed below, go into detail about aspects for documentation, there are several authors that have written about graffiti art that have been extremely useful to support this research and are often cited across various disciplines of study. Ross (2016) provides a glossary of graffiti-related terminology that includes several definitions for types and styles of graffiti (475-79). Austin (2001) focuses on New York City and the graffiti wars among graffiti writers, and eventually between writers and law enforcement, resulting in the city-wide campaign to buff the subway trains of all graffiti under the auspices of mayor Ed Koch in the late 1970s and 1980s. Austin provides a rich history of the art form, as well as the sociological and political underpinnings and responses to what many viewed simply as vandalism.

Riggle (2010) addresses graffiti and street art from an arts perspective and eloquently forefronts the importance of the street itself in the production and reception of the works. Waclawek (2011) explores the evolution of graffiti from its modern beginnings to the appearance of street art as a secondary art form. Her writing explores the differences in these two terms and the overlaps they engender. Both graffiti and street art share a history and many commonalities, and the labels are very often used interchangeably by those only casually familiar with the art forms. Austin (2010) further articulates a difference between graffiti and graffiti art while MacDowall (2019) highlights the often antagonistic relationship between graffiti artists and street artists. Schacter (2017) describes street art as a distinct art period, marked by stylistic conventions and classificatory permutations in what has been referred to as the post-graffiti era of the late 1990s and early 2000s. Those within the respective communities will be quick to point out the differences between graffiti and street art, but when examining hundreds of widely distributed and contributed image galleries online, the distinction goes beyond the scope of this research. The image websites explored herein have shared a mixture of graffiti and street art.

Very little research has explored the use of specific controlled vocabularies for use with graffiti art collections, though some authors mention terminology or typologies for the works in their writing. Biaggini (2018) investigated visual interventions, or wall painting, in Buenos Aires and formed a very broad set of categories for the works, including muralism, stencils, tribute works to those who had died, graffiti, educational murals, and works relating to the war...
of the Malvinas Islands. These reflect categories found in the current research, as well as categories specific to the location studied. MacDowall (2019) excavated the top hashtags relating to graffiti on the social media site Instagram, though these were only terms applied to graffiti in general, a typology of terms that could be taken to mean graffiti. He notes coherence with the term street art that is not afforded to graffiti, the latter which has numerous variations in use around the world on Instagram (graffiti, graff, instagraff, etc.) (30). Gottlieb (2008) approached graffiti from a library and information science perspective for the purposes of developing a classification that could be used with collections of graffiti art images. Her work is based upon the theories of Erwin Panofsky and interviews with several graffiti artists, used to gain insight into various named aspects of graffiti form. Her graffiti styles classification expresses names for fourteen styles and aspects such as legibility, number of colors used, symmetry, letter overlap, outlines, and fills. While interesting as an example of classification, it is not known whether this system has ever been applied to a collection of graffiti images. It remains an example of research into graffiti art styles and dimensions of the artworks for a generalist art audience but does not focus on the particular aspects of documentation used within the graffiti art community itself.

Within library and information science, and KO in particular, there have been several studies of user communities and how they each address organization of collections and choices of terminology for description. Among these, Lee and Trace (2009) researched those who collect rubber ducks, revealing categories of information for aspects such as materials, value specifications, series, years, companies, and countries. Hartel (2010) examined the information organization behaviors of hobbyist gourmet cooks and revealed facets of entities used to organize information, features, spaces, and processes. Cho et al. (2018) discussed the need for distinct anime genre facets, identifying nine facets and 153 individual terms for use in describing anime genres. Summarizing research on the online image platform Flickr, Stuart (2019) discusses how images are organized by the site using metadata embedded within images as well as external organization by the user application of hashtags, the addition of titles, and descriptive text. Graf (2016) applied domain analysis to graffiti zines from the 1990s and 2000s to reveal process and product terminology used within the graffiti art community, then compared the resultant term list with available controlled vocabulary within the Getty Research Institute’s AAT. At the time, there was only a 15% match between the top twenty community terms and AAT vocabulary, though by 2018 that percentage had jumped to 70% for the same list of terms after the Getty added thirteen more of the original twenty terms to the thesaurus. This in and of itself indicates the popularity of graffiti art and the value of terminological studies to highlight specific art community descriptive practice.

Graf (2018, 2020) expanded upon this research and augmented the findings with the results of interviews with graffiti website curators. The works of Graf, Cho et al., and Stuart, like that reported herein, explore what named aspects are already in use for organization and make distinctions among them. Benedetti (2000), in writing about vocabulary used to describe folk art, emphasized the importance of allowing those on the edges of the institutional art world to describe their works using their own vernacular, which helps engender trust between what may be considered art insiders and outsiders. This is a concept easily transferrable from folk art to graffiti and street art, also on the fringes of the traditional art world.

Alongside the benefit of nurturing trust from marginalized or otherwise non-institutional communities of practice is the acknowledgement that descriptive practices and vocabularies are often not known, not researched, overlooked, or otherwise foregone in favor of defaulting to controlled vocabularies already in common use. This discounts what Terras (2010) refers to as intuitive metadata, which she explored in online collections that were curated by non-professionals. The website collections in her study share challenges also inherent in graffiti image collections. Amateurs in her study were those curating website collections not associated with formal institutions and organizing images of objects not always owned by the curators, forcing them to rely at times on data supplied by contributors. Such data are often incomplete, unreliable, unverifiable, or entirely absent, which renders description challenging. This mirrors the experience of numerous graffiti art image galleries in this study, many of which rely on contributed images to fill out their collections. The excavation of ontological formation on this and similar websites can be used to guide the development of useful website architecture, metadata schemas, vocabularies, and other KOSs to guide users of such websites in their search for and use of information (Srinivasan 2005). Further, such research adds to the discussion of value placed upon professional curation over and above what Dallas (2016) refers to as curation in the wild, and the ongoing need to address the relationships among professionals and non-professionals (the latter who often have the most intimate experience with their collections), controlled vocabularies and fluid, organic ontologies, and with curators of information and users of information.

1.2 Methodology

Despite a lack of institutional documentation and organization of graffiti art images, a plethora of online galleries exist that are managed by independent groups and individuals around the globe. To choose a selection of these galleries for
examination, one of the earliest, largest, and most popular websites for graffiti art images was chosen as a source for all other websites in the study. Art Crimes (graffiti.org) was started in 1994 and “was the first graffiti site on the net” according to their about page information (https://www.graffiti.org/index/story.html). The site, a volunteer and collaborative effort, gathers and shares information about graffiti and graffiti art as well as curates a large image collection and links to hundreds of other graffiti websites. The list of website links from Art Crimes became the source of all the websites used as a sample in this research, as well as an individual site included in the overall list of websites. It should be noted that the words website and site are used interchangeably herein.

In 2017, Art Crimes included a list of 709 links to other graffiti-related websites. Each of these websites was visited to determine appropriateness for the study. Sites that were dead, i.e., no longer available, used navigation labels and main text solely in a language other than English, were completely commercial (such as sites dedicated to an artist’s professional work for sale), were only links to a social media site such as Instagram or Flickr, or were otherwise not sharing image collections of graffiti and graffiti art were excluded. One link was to a subpage of Art Crimes itself and was, therefore, excluded as well. Social media sites such as Instagram and Flickr were excluded from the study because of the structures these platforms impose on the size of images (Instagram) and on the ability to organize images, galleries, and sub-galleries (Instagram, Flickr). Once this initial examination of all 709 links was complete, 241 websites remained for inclusion in the study. Table 1 provides a breakdown of all sites examined for inclusion.

<table>
<thead>
<tr>
<th>Websites</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>241</td>
<td>Live sites included for study</td>
</tr>
<tr>
<td>318</td>
<td>Dead, empty, or moved</td>
</tr>
<tr>
<td>64</td>
<td>Other languages</td>
</tr>
<tr>
<td>57</td>
<td>Artist’s professional site (not graffiti or street art)</td>
</tr>
<tr>
<td>20</td>
<td>Not relevant (music group, advertising, etc.)</td>
</tr>
<tr>
<td>8</td>
<td>Other social media only (Flickr, Instagram, etc.)</td>
</tr>
<tr>
<td>1</td>
<td>Art Crimes sub-page, not website</td>
</tr>
</tbody>
</table>

Table 1. All websites examined for inclusion in the study.

Each of the 241 websites was then examined for explanatory text and navigation label text. Relevant text from the home page of each website, as well as any text from an about page or a page speaking about the history and/or the purpose of the website, was manually harvested and copied into an Excel file. Navigation labels were usually found on the home page of the website, but further exploration sometimes revealed deeper structure with subpage labels, most often for more specific division of image galleries. The labels given to pages on the websites are indicative of aspects for organizing images of graffiti and graffiti art within this study and were also copied into the Excel file.

Once all of the data described above were collected, they were imported into QDA Miner software for qualitative coding. As the coding progressed, evidence of two top level categories appeared, one for description of the websites themselves, and the other for description of the work images featured on the websites. After all data were coded, the category relating to the websites themselves was divided into two sections, one for general website description and the other for links to other media, such as associated Instagram, Flickr, or Facebook feeds. This category and its sections will be described briefly. The other category consists of information about the organization of the work images themselves, and is the main focus of this report.

2.0 Results

2.1 The categories

The coding developed organically, soon evidencing two main categories of navigation labels: those that describe aspects of the websites themselves and those that are used to organize image galleries of graffiti art, representing the actual works. This is the first main division in the complete list of codes. It is important to note that within the latter category, images of works are examined and not the actual works themselves. This is an important distinction to make and impacts the creation and application of metadata. In formal systems for art documentation, such as the Visual Resources Association’s Cataloging Cultural Objects data content standard, clear distinctions exist between a work and an image of a work and how to link records of each together while avoiding confusion (Baca et al 2006).

While the focus of the research is to discover which aspects of the works are being used as divisions for organization of the image galleries, the data come from the navigation labels, some of which do not speak directly to the works or their images. These, therefore, were not simply ignored but were gathered into their own separate category and further subdivided into the two sections that will be discussed briefly. The larger category of codes, those that relate to aspects of the works as reflected in their images, remain the focus of the work reported herein.

To make the structure of the following results easier to understand, the coding scheme is summarized in graphic form in Figure 1. Each of the two large categories of codes are further divided into sections, which in turn are subdivided into a series of codes representing aspects, and herein referred to as facets. It is noted that the term “facets” will be used in this research in a descriptive sense to denote aspects
or groupings and is not intended, at this time, to reflect parts of a faceted classification.

2.1.1 The Websites category

The Websites category pertains to navigation labels not directly about the image galleries or works but rather about the sites themselves. This includes information about what is offered on each site apart from images and includes about pages, contact information, shopping, forums, associated media, and videos on various topics. There are two sections to this category, one for general website information and one for links to other media. The complete breakdown of this website category and these two sections can be seen in Table 2. The table indicates the codes applied to each section of the websites category and these two sections can be seen in Table 2. The table indicates the codes applied to each section of the websites category, as well as how many times individual codes were applied across all 241 websites, what percent this count represents of all codes applied, the number of sites that employed this type of navigation label, and the percent of all websites that employed this type of navigation label. The tables in this report are each arranged similarly, listing individual codes (also referred to as facets) in descending order from most sites to least sites warranting each.

The two most commonly encountered sections of the Website category are for contacting the website curators (“contact”) and for reading more about the site itself (“about”). Visitors to the websites can also often shop for merchandise that either advertises or supports the sites, view videos of interviews with artists or read transcripts of interviews, read disclaimers from the website curators (often pertaining to the illegality of some of the works featured on the site), contribute graffiti art images to the site (commonly referred to as flix), or participate in forums, ask questions, learn about how to create graffiti art of various types, and subscribe to updates posted to the site. A limited number of websites featured a user poll, soliciting opinions from visitors, and even fewer sites, only two out of the 241, offered a glossary of terms. The OtherMedia section indicates when a website offered a place to list links to other related websites, or links to associated social media accounts, a blog, or a published book for sale pertaining to the site and its activities. It was common for a site to have a place to provide links to other graffiti art websites or online outlets where graffiti art supplies can be purchased. Many sites provided links to associated social media sites, with Facebook and Instagram the most commonly referenced. While this category describing the websites does not speak directly to the works in image galleries, it provides valuable information that can be used to further illuminate the commonalities across graffiti art image collections, their goals, users, online presence, and organization.

2.1.2 The Works category

The other overarching category of codes applied to the data is that related to the organization of graffiti art image galleries, or to the works themselves as represented by their im-

Figure 1. Coding hierarchy structure, indicating two categories and their individual sections.
ages. This large category was divided into four sections: General, Support, Type, and Location. Each of these sections is further subdivided into specific codes for facets of the works that are used on the websites to organize images. Each of the four sections will be discussed in detail.

### 2.1.2.1 Work General section

The Work General section is subdivided into facets for organization that describe features of the works that did not organically develop into separate groups. One could argue for an additional section on time-based codes, such as Year, Month, Day, and Decade, but these were not very heavily warranted so they remain in the General section with other adjectival codes. The application of the Work General codes is shown below in Table 3.

The name of an artist is the most common way that works are organized among the General Work facets. This code was applied 14,439 times across all of the 241 sites and represents 71.2% of all codes used in this study. No other type of code for graffiti artwork galleries was warranted anywhere near this many times so it is easily seen as a very important way to organize graffiti artworks. The second most commonly assigned code to organize works is the city code, a facet of the Location category, with 1,637 applications in this study. No other codes were applied over 1,000 times.

This type of code was applied to navigation labels whenever the name of an artist or crew was used as a link to a gallery of works by that artist or crew. An example from the websites where this code was liberally applied can be seen in Figure 2.

In this gallery page screenshot from website 50mm Los Angeles, there is a list of numerous links to individual gallery pages where images are organized according to criteria represented by the links themselves. The majority of these links listed in the alphabetized paragraph are the names of artists. Interspersed among the artist names are the names of cities (Amsterdam Graffiti), US states (Arizona), countries (Brazilian graffiti, Canadian graffiti), crews (C.U.L.T. Crew, AWR Crew), supports (billboards, boom boxes, bus,

<table>
<thead>
<tr>
<th>Sites</th>
<th>Count</th>
<th>% of Codes</th>
<th>Sites</th>
<th>% of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contact</td>
<td>147</td>
<td>0.7</td>
<td>133</td>
<td>55.2</td>
</tr>
<tr>
<td>About</td>
<td>145</td>
<td>0.7</td>
<td>110</td>
<td>45.6</td>
</tr>
<tr>
<td>Shop</td>
<td>77</td>
<td>0.4</td>
<td>60</td>
<td>24.9</td>
</tr>
<tr>
<td>Videos</td>
<td>61</td>
<td>0.3</td>
<td>43</td>
<td>17.3</td>
</tr>
<tr>
<td>Disclaimer</td>
<td>33</td>
<td>0.2</td>
<td>26</td>
<td>10.8</td>
</tr>
<tr>
<td>Interviews</td>
<td>51</td>
<td>0.3</td>
<td>22</td>
<td>9.1</td>
</tr>
<tr>
<td>ContributeFlix</td>
<td>18</td>
<td>0.1</td>
<td>16</td>
<td>6.6</td>
</tr>
<tr>
<td>Guestbook</td>
<td>21</td>
<td>0.1</td>
<td>15</td>
<td>6.2</td>
</tr>
<tr>
<td>History</td>
<td>18</td>
<td>0.1</td>
<td>9</td>
<td>3.7</td>
</tr>
<tr>
<td>Map</td>
<td>10</td>
<td>0.0</td>
<td>9</td>
<td>3.7</td>
</tr>
<tr>
<td>MyAccount</td>
<td>13</td>
<td>0.1</td>
<td>9</td>
<td>3.7</td>
</tr>
<tr>
<td>Forum</td>
<td>9</td>
<td>0.0</td>
<td>8</td>
<td>3.3</td>
</tr>
<tr>
<td>FAQ</td>
<td>7</td>
<td>0.0</td>
<td>7</td>
<td>2.9</td>
</tr>
<tr>
<td>HowTo</td>
<td>76</td>
<td>0.4</td>
<td>7</td>
<td>2.9</td>
</tr>
<tr>
<td>Subscribe</td>
<td>6</td>
<td>0.0</td>
<td>6</td>
<td>2.5</td>
</tr>
<tr>
<td>Poll</td>
<td>3</td>
<td>0.0</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td>Glossary</td>
<td>6</td>
<td>0.0</td>
<td>2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OtherMedia</th>
<th>Count</th>
<th>% of Codes</th>
<th>Sites</th>
<th>% of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Links</td>
<td>107</td>
<td>0.5</td>
<td>90</td>
<td>37.3</td>
</tr>
<tr>
<td>Blog</td>
<td>50</td>
<td>0.2</td>
<td>40</td>
<td>16.6</td>
</tr>
<tr>
<td>Facebook</td>
<td>12</td>
<td>0.1</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>Instagram</td>
<td>11</td>
<td>0.1</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>Twitter</td>
<td>9</td>
<td>0.0</td>
<td>7</td>
<td>2.9</td>
</tr>
<tr>
<td>Flickr</td>
<td>8</td>
<td>0.0</td>
<td>6</td>
<td>2.5</td>
</tr>
<tr>
<td>Book</td>
<td>8</td>
<td>0.0</td>
<td>6</td>
<td>2.5</td>
</tr>
<tr>
<td>YouTube</td>
<td>4</td>
<td>0.0</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td>Pinterest</td>
<td>3</td>
<td>0.0</td>
<td>2</td>
<td>0.8</td>
</tr>
<tr>
<td>Tumblr</td>
<td>2</td>
<td>0.0</td>
<td>2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 2. Website category, sections, and codes/facets.
### Table 3. Work General codes/facets.

<table>
<thead>
<tr>
<th>General Facets</th>
<th>Count</th>
<th>% of Codes</th>
<th>Sites</th>
<th>% of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artist</td>
<td>14439</td>
<td>71.2</td>
<td>50</td>
<td>20.7</td>
</tr>
<tr>
<td>Event</td>
<td>89</td>
<td>0.4</td>
<td>31</td>
<td>12.9</td>
</tr>
<tr>
<td>Gallery</td>
<td>49</td>
<td>0.2</td>
<td>29</td>
<td>12.0</td>
</tr>
<tr>
<td>Year</td>
<td>227</td>
<td>1.1</td>
<td>27</td>
<td>11.2</td>
</tr>
<tr>
<td>New</td>
<td>35</td>
<td>0.2</td>
<td>26</td>
<td>10.8</td>
</tr>
<tr>
<td>Old</td>
<td>35</td>
<td>0.2</td>
<td>26</td>
<td>10.8</td>
</tr>
<tr>
<td>Featured</td>
<td>27</td>
<td>0.1</td>
<td>20</td>
<td>8.3</td>
</tr>
<tr>
<td>Inside</td>
<td>11</td>
<td>0.1</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>RIP</td>
<td>75</td>
<td>0.4</td>
<td>10</td>
<td>4.1</td>
</tr>
<tr>
<td>RatedHigh</td>
<td>14</td>
<td>0.1</td>
<td>8</td>
<td>3.3</td>
</tr>
<tr>
<td>Legal</td>
<td>15</td>
<td>0.1</td>
<td>7</td>
<td>2.9</td>
</tr>
<tr>
<td>Outside</td>
<td>7</td>
<td>0.0</td>
<td>7</td>
<td>2.9</td>
</tr>
<tr>
<td>Month</td>
<td>35</td>
<td>0.2</td>
<td>5</td>
<td>2.1</td>
</tr>
<tr>
<td>Color</td>
<td>12</td>
<td>0.1</td>
<td>4</td>
<td>1.7</td>
</tr>
<tr>
<td>Day</td>
<td>5</td>
<td>0.0</td>
<td>4</td>
<td>1.7</td>
</tr>
<tr>
<td>Decade</td>
<td>8</td>
<td>0.0</td>
<td>4</td>
<td>1.7</td>
</tr>
<tr>
<td>Illegal</td>
<td>5</td>
<td>0.0</td>
<td>2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Figure 2. 50mm Los Angeles gallery page (http://www.50mmlosangeles.com/gallery.php).
canvas), RIP memorials (Amore (r.i.p.), Ayer (r.i.p.)), and events (50mm Production 01.16.05), among other types of facets. The screenshot only shows through the beginning of the letter “c,” but on the live site continues all the way through the alphabet down the page. This image also illustrates the types of navigation labels that were collected and coded during the research. It can be seen that while fifty sites used this type of facet for organization of graffiti artworks, it was used by some sites several hundred or more times for various artists. The bigger the website, the greater the number of artists and works represented. Some sites are dedicated to works by only one artist or one group of artists, referred to as a crew. For sites focusing on one artist, it would not make sense to use the name of the artist as an organizational tool, though sites focusing on an individual crew do use the members’ names to indicate when an individual created a work or was responsible for the design of a crew work.

Named events are also a way that graffiti image galleries are organized. Images from annual or one-off festivals are often gathered together on a page of their own. These events can be important markers in the evolution of the art form that bring together disparate artists to work near each other or together on pieces, share styles and techniques, and gain knowledge and insight into local cultures. Smaller events often highlight local styles and are then shared globally in event galleries. The Gallery code was applied when a site displayed works from or featured in a formal gallery show. These are sometimes gatherings of works by one artist, or many, sometimes on canvas or inside walls, and sometimes featuring works on the streets, yet curated through sponsorship by a gallery.

Going down the list, works are often organized by year, most often on sites that are dedicated to the works of one artist or crew. This provides a valuable view of an artist’s development over time as easy comparisons can be made between early and more recent works. Codes related to the popular Year facet include Month, Day, and Decade, appearing near the bottom of this section list. The Day code refers to images taken or works created on a particular day, and could at times be related to the Event code, but inspection of individual images would be necessary to differentiate the two. This code was only warranted on four of the 241 sites. The New code was applied when a website had a specific image gallery for works new to the site. This might include things that were produced long ago, but recently submitted or added, brand new works, or anything in between. It does not refer with any consistency to the age of the works themselves. The Old code, on the other hand, more often referred to works that were indeed older, painted earlier in the history of modern graffiti art. These are often used as historical markers, highlighting pioneers in the art form and paying homage to their individual styles. The Featured code has some semantic overlap with the Old code, but is more of a curated look at a particular artist or crew. In image two, the navigation label “L.A. Legends” can be seen on the website banner for 50mm Los Angeles. This indicates where the Featured code would be applied. This image gallery features works by known L.A. artists and is divided into sections for each, with images and transcribed interviews.

The Inside code was used to represent galleries showing works committed inside buildings or inside other structures. This is sometimes used to differentiate works that are done outside on walls, billboards, and other public-facing structures. The complementary code Outside is used to indicate the opposite, though it was not warranted as often as the Inside code. This is not surprising as graffiti is commonly encountered outside by default. This occurs with the Legal and Illegal codes as well. Most graffiti is explicitly defined as illegal, which becomes the default. The Legal code was warranted three times more often than the Illegal code for this reason. In the graffiti community, legally created works do not carry the same kind of respect as illegal ones. It is much easier to create a beautiful, intricate legal piece because of the ability to work without pressure, in daylight, and sometimes even with financial support. A beautiful and intricate illegal piece garners a lot of admiration from other graffiti artists because of the constraints placed upon the artist to work under pressure, often quickly, and often in the dark and in difficult locations. Using a code such as Legal (or Illegal) is adding important information regarding the creation of the works.

The RIP code was applied to galleries dedicated to particular artists who have died. This may include works by the deceased as well as works by others in their memory or honor. It is common to add the initials RIP after an artist’s name when they die. In this way, some navigational labels were coded for an artist name as well as for RIP. Further research could differentiate between works that are by deceased artists and those created in honor or memory of deceased artists. The Rated High code was only warranted on eight websites in this study. To use this type of code, a website had to offer some kind of user ranking system, either all the time or during specific open periods on the site. Users can submit their favorites in this way and the website can choose to use some kind of cut-off in ranking to determine which works can be gathered for this purpose. The Color code was only used on four sites and does not represent a popular way to organize works.

### 2.1.2.2 Work Support section

The second section of the Work category is that for various kinds of supports upon which graffiti art is created or placed. Sometimes works are created in situ, while other times they may be created in a studio or elsewhere and moved to another site and mounted, glued, screwed, sewed, or otherwise affixed.
Three facets of support stood out in this section: Canvas, Walls, and Trains. The Canvas support is interesting, because it indicates the importance of noting when works are created in a studio and not on the street. Graffiti is often defined by its use of the street, broadly conceived (Riggle 2010) so that graffiti works created in a studio, on canvas, could be perceived as graffiti-style instead of graffiti itself. This code encompasses works created in a studio or displayed in a gallery, as opposed to found in the more traditional graffiti context. Terminology used to represent navigation labels coded as Canvas include painting, art, prints, studio, color works on paper, commission, portraits, watercolor, and street sellout art.

The next Support facet for Walls is used by sixty-five sites, indeed a very common support upon which graffiti is found. The third most commonly warranted Support code is for Trains, used on fifty-one sites but used more often on those fifty-one sites than any other Support facet. This means that a site may have a gallery area for trains that is further subdivided so more than one use of the Train code is warranted. Various named kinds of train graffiti will be seen in the next section on Types.

Trains represent a complex Support facet in this study, which begs further explanation. In this list of Supports, one can see Trains, Freights, and Subway Cars. All of these could be called Trains, though there are subtle differences. The word Train is the most generic of the three. Freights are a specific type of train with flat sides and no windows, carrying not passengers but things. They represent a specialized type of graffiti surface that has long been exclusively favored by certain artists. The freight car offers a moving support that can take an artist’s work across entire countries or continents. Styles have moved from coast to coast in the United States and other countries in this way. Individual freight cars are recombined in train yards for new journeys, and works may be seen in different juxtapositions over time and space. Subway cars, on the other hand, are very different. They carry people and usually have windows to break up the sides. They are also notably found in larger urban areas and do not usually go far beyond individual city hubs. Subway cars are often much more difficult to access than freight cars and more difficult supports upon which to create large and complex pieces. They are also cleaned, or buffed of graffiti, frequently (see Austin 2001). For the purposes of this study, navigation labels that indicated galleries for works on trains were coded as such, but if a site explicitly used labels for subway cars or for freights, those codes were applied. The Subway code was applied when works were on the subway structures but not referring to the subway cars themselves. It is noted that all of these navigational terms used on the sites may have overlap. Further study would be needed involving investigation of individual images to make a more precise reporting of these Supports, which goes beyond the scope of this research. It is sufficient to say that graffiti on trains of various types represents one of the most commonly cited Supports in this study.

There is a similar issue with the Subways and Tunnels facets in this section. Tunnels may be subway tunnels, or they may simply be other tunnel structures. Graffiti is commonly found in subways, inside the tunnels and also outside.
on walls or other structures along the tracks. This possible overlap is acknowledged, but when a site indicated navigation labels specifically for subways, this code was applied. Tunnels, when similarly stated, were coded as Tunnels. Further inspection of individual images in these galleries would be required to make a more precise reporting of these Supports, which goes beyond the scope of this research.

The artist sketchbook, referred to in the graffiti art community as a “black book” or “blackbook,” is another notable Support in this section. The Getty Research Institute includes the term in the Art and Architecture Thesaurus (AAT) as “black book (graffiti)” and notes (Getty Research Institute 2004) the term as “sketchbooks in which graffiti artists develop ideas, practice tags, and acquire other artists’ signatures.” Twenty of the study websites reserved individual image galleries for works from artists’ blackbooks.

Besides trains, other moving vehicles used as supports were either grouped as Cars, Trucks, and Vans together, or separately as Buses. The Body support refers to either body painting or tattoos of graffiti-style artwork. Billboards and Signs are similar, but Signs refers to smaller surfaces such as municipal or traffic signs. Shutters refers to pull down metal or wood doors facing the street used to cover small shops when closed.

### 2.1.2.3 Work Type section

The Work Type section includes descriptive terminology used to separate individual image galleries dedicated to various types or styles of graffiti art. This section represents the largest of the code groupings and includes a broad range of style vocabulary. Detailed explanations of what each of these facet terms represent can be found in Graf (2018). Discussion here will focus on broader themes present within this section and relationships among included facets.

Nearly a quarter of all sites dedicated a part of their website to Sketches. This is interesting, because sketches represent ideas or layouts that are not actually completed graffiti artworks. The more complex a planned piece, the more time required to complete it and the more important it is to draw up what the final work might look like so that precious time on site can be better utilized. This code was applied when encountering navigation labels including words like sketches, drawings, outlines, illustrations, and phrases using these words, such as charcoal sketches, ink outlines, and sketch battles. Because a graffiti artist’s blackbook is often used to create these sketches, there is overlap between these two facets of their respective categories.

The second most warranted type facet in this category is for Graffiti. This may seem so obvious as to not be useful, but it was used by roughly 20% of all sites in the study. Variations in the terminology that earned this code include graffiti, graff, vandalism, and assorted sub-galleries of named geographic facets (Amsterdam Graffiti, LA Graffiti, etc.). Each of these galleries was visited to make sure that the focus was indeed on graffiti, especially when terms like vandalism had been used. The next most popular type facet was Other. This code represents graffiti types that do not fit into other facets used as types, and were only used once or by one site. The Other code was applied when the word “other” was used to set aside images (other graffiti pictures, other works) or when a label was used that did not have any obvious meaning. Examples of this include when a site simply numbered galleries, or had place-saving text such as “empty,” as if the website administrator had simply not given a name to a gallery so the website template term was all that was used. Other examples of text that earned the Other code in this study include but are not limited to, scratchiti, motorcycles, mixed media, screenprints, scenes, wallpapers, cardboards, unusual spots, bunker, burners, quickies, planes, and trams.

The next three codes in this category, CommercialDesign, StreetArt, and Murals, share an aspect that is important to this study. All of these imply work that is legal and possibly done for profit. The terms street art and murals have been used interchangeably at times in the literature, but common practice is to use these terms as distinct from graffiti. Commercial design implies a work created for profit at the request of a customer. While commercial designs, street art, and murals can and are often done in graffiti art styles, the association with these descriptive terms separates them from traditionally created graffiti art. They are often very large and beautifully executed pieces due to their creation under legal and often sponsored circumstances, and they are afforded protections that most graffiti art is not. They are documented and shared widely, as this research indicates, placing them near the top of terms for organization used in this category. The most important aspect of note is the use of terminology for them that separates them from traditional graffiti, speaking more to the environment surrounding their creation than to a specific style. The research does show that the CommercialDesign code was used almost exclusively by websites curated by individual artists, graffiti crews, or other artist collectives. As opposed to sites that are curated to feature a broad range of works from many artists, these sites display an intrinsic motivation for profit from exposure to their work and gaining further commissioned work.

Codes used for types of train graffiti in this category include Wholecars, EtOEs, TtOBs, and TrainPanels. Note that other Types included here may be found on trains, but these four codes refer to the sizes of works that are done specifically on trains. This type of granularity was understandably found most often on sites that featured numerous works on trains and provides a way to divide up the organization of such galleries. A work can cover an entire train car, or a train...
car from end to end (EtoE), but not going all the way from top to bottom. Conversely, a work can cover a train car from top to bottom (T toB), but not from end to end. These distinctions are important to the graffiti artist, because they indicate various levels of difficulty and time required to complete an illegal work. Train cars are tall and require the use of ladders and equipment to reach the top for a T toB piece. An EtoE may, therefore, be considered difficult in that it takes time but not as difficult as painting a T toB and not as difficult, and, therefore, respected if well done, as completing a Wholecar work. TrainPanel is used for a smaller part of a train than the other train-related terms here. Again, there may be some overlap in the actual types of works featured in these image galleries.

### 2.1.2.4 Work Location section

The last of the Work category sections is that for geographic location of works. The codes applied within this section are shown in Table 6. This type of information was included on the sites using standard textual labels, some of which can be seen earlier in Figure 2 but sometimes in use on graphic maps where users can click on different parts of the world or a specific country and see works from those locations.
Nine sites within the study used some form of map for image gallery navigation. An example of a navigational map is shown in Figure 3.

Whether the geographic locations were listed as standard text or as labels on an interactive map, each link to an individual gallery of images for a specific location was coded appropriately using facets of this section. Cities and countries were by far the most commonly used geographic division across all websites. Forty-three percent of all sites used some kind of organizational feature by individual cities and 37% by countries. There are several factors at play regarding the use of geographic labels for organizing graffiti art images, not the least of which is the focus of the site itself and the way that images are gathered for inclusion. Many websites specifically focus on graffiti art from a named country (Australian Graffiti, Irish Street Art, etc.), or city (Bristol Street Art, The Helsinki Connection, Miami Graffiti, etc.) and, therefore, exhibit a narrower geographic focus than sites that seek to highlight works from around the world, often soliciting images from global visitors to the sites (12 Oz. Prophet, Can Control, Art Crimes, FatCap, etc.). The focus of the site and the number of images to organize unsurprisingly contributes to the need for different types of geographic facets. Dividing by continents only makes sense if you have a global image collection. Division by specific landmarks or parts of cities is useful with a collection that includes numerous works from one metropolitan area, such as New York or London.

The CountryParts code was applied when large countries were divided by logical divisions, such as the United States into East Coast, West Coast, or Midwest. The States code was applied to not only US states, but Canadian provinences and Australian states. CityParts refer to named areas or neighborhoods used as individual galleries, for example when listing Tenderloin, Castro, or Mission in San Francisco, or Plaistow, Tottenham, or Crouch End in London. The SpecificLandmarks code was applied to named buildings, parks, train yards, or other such individual locations. These named locations are often popular spots for graffiti art and have earned a reputation as such.

<table>
<thead>
<tr>
<th>Location Facets</th>
<th>Count</th>
<th>% of Codes</th>
<th>Sites</th>
<th>% of Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cities</td>
<td>1637</td>
<td>8.6</td>
<td>43</td>
<td>17.8</td>
</tr>
<tr>
<td>Countries</td>
<td>543</td>
<td>2.8</td>
<td>37</td>
<td>15.8</td>
</tr>
<tr>
<td>SpecificLandmarks</td>
<td>73</td>
<td>0.4</td>
<td>13</td>
<td>5.8</td>
</tr>
<tr>
<td>CityParts</td>
<td>94</td>
<td>0.5</td>
<td>13</td>
<td>5.4</td>
</tr>
<tr>
<td>World</td>
<td>22</td>
<td>0.1</td>
<td>12</td>
<td>5.0</td>
</tr>
<tr>
<td>Continents</td>
<td>42</td>
<td>0.2</td>
<td>11</td>
<td>4.6</td>
</tr>
<tr>
<td>States</td>
<td>117</td>
<td>0.6</td>
<td>6</td>
<td>2.5</td>
</tr>
<tr>
<td>CountryParts</td>
<td>10</td>
<td>0.0</td>
<td>5</td>
<td>2.1</td>
</tr>
<tr>
<td>Address</td>
<td>2</td>
<td>0.0</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>Intersection</td>
<td>27</td>
<td>0.1</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>Undisclosed</td>
<td>2</td>
<td>0.0</td>
<td>1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

*Table 6. Work Location codes/facets.*

*Figure 3. World map from site Bombing Science (https://www.bombingscience.com/graffiti-map/).*
It is notable that no information more specific than a part of a city or named landmark was used for organization of graffiti art images in this study. Only one site had navigation labels referring to a specific address. A street intersection was used only once, and the code Undisclosed was applied only once, the latter representing that works were specifically not located, their location either not provided, explicitly withheld, or otherwise unknown. Further research would be required to determine if GPS coordinates for works were ever provided for individual graffiti art images, but this is beyond the scope of this research. Anecdotally, GPS information was never encountered for images during the website examinations.

3.0 Summary

This study has explored the knowledge organization practices of graffiti art website curators from 241 sites hosted around the world using a domain-analytic approach. All navigational text used on the sites was coded into two broad categories, one for description of aspects of the websites themselves, and one for organization of works. The Websites category was further divided into sections specifically relating to the site or to social media links. The Works category was further divided into sections labeled as General, Support, Type, and Location. Each of these sections were further subdivided into individual codes or facets and were described herein as the focus of this report.

Even the most commonly used facets for organization of graffiti images in this study rarely were used by more than 20% of all 241 sites. Those that were include the General Artist code (20.7% of all sites), the Support Canvas code (32%), the Support Walls code (27%), the Type Sketches code (23.2%), and the Type Graffiti code (20.7%). Variety among the types of codes across all sites was influenced by a number of factors. Larger sites with more images displayed warranted more granularity in organization and earned a broader range of codes for things like General work facets, work Support facets, and work Type facets. The geographic focus of each site impacts the range of work Location facets. Sites with a specific focus, such as those covering works on trains, used facets for description with more granularity in those areas.

Internal and external motivation were coded during the study and were shown to correlate with terminological choices for description and organization. Stated motivation was found on 112 of the 241 sites. Of these, seventy-two sites were coded as internally motivated and forty as externally motivated. Internally motivated sites, those run by individual artists or crews seeking exposure with an explicit intent to garner commissions and sales of artwork, tended to avoid using terms that were more common in the graffiti art community, excluding words like graffiti and vandalism, and favoring instead terms including street art, mural, urban art, and vocabulary that is not associated with illegal activity. Externally motivated sites, those not explicitly seeking paid work or commissions, but rather sharing artworks for enjoyment and study, did not shy away from graffiti terminology. In fact, the externally motivated sites displayed the broadest range of facets for the organization of image galleries while internally motivated sites used the widest range of art style vocabulary (see Graf, 2018).

4.0 Conclusion and further research

This domain is ripe for further research on many fronts. There is a group of websites from the original 241 examined herein that host very large image collections from various artists around the world. Research on this subset of large, externally motivated sites may prove valuable to refine this study’s results with information on only the largest, broadest, and most descriptive collections, and those not displaying possible biases in terminology fueled by an underlying motivation to sell works. This research is in process by the author.

There are also many other opportunities to refine and expand upon this type of domain analytic work to increase understanding of the knowledge organization structures in use within the graffiti art community. Work could be carried out to examine instantiation of individual works over time and space and how they are documented by various agents. Because images for the aforementioned larger sites are often solicited from and received by a wide range of individuals, differences in documentation practices involving angles, contextual inclusion or exclusion, instantiation over time displaying changes in condition of works, and a myriad other variations may be revealed. Related to instantiation research on individual works is research on individual artists over time and space. These types of research require more refined focus on individual images and their textual accompaniments.

There is also a wide range of research that can be conducted on social media sites such as Instagram, arguably the largest image sharing platform in use around the world. The organizational practices of Instagram users are largely shaped by the structure and limitations of the platform itself. Analysis of image tagging practices applied to graffiti works on Instagram could lead to further precision of style, location, and type facets already found in this examination. Research of this type is also in development by the author.

Graf (2016) excavated graffiti art terminology from a series of graffiti zines from the 1980s to 2000s and compared it with what was available in the AAT. Only three of the top twenty occurring graffiti terms from the zines were found in the AAT in 2016, but within two years the AAT had added an additional eleven of these twenty terms, bringing the in-
clusion of studied terms from 15% to 70%. This was an important addition to one of the most popular controlled vocabularies for art-related terms in use around the world. There are further terms that could be added to the AAT, but it is also acknowledged that those doing the bulk of the documentation, description, and organization of graffiti artworks online, the aforementioned graffiti art community, are not using such tools with their image collections.

This research may inform not only a taxonomy of graffiti art terms and thesaural vocabulary but also awareness of facets for description that may be useful for the graffiti art community if made available as a metadata schema or application profile. It may also be useful for libraries, archives, museums, and other public and community art associations as more traditional institutions consider curating online collections of graffiti art and street art images themselves and seek to respect the terminological distinctions applied by those within the graffiti art community. The work remains an example of the use of domain analytic techniques engaged in the discovery of organizational practices of an extra-institutional user community, a domain, and the methodology has demonstrated the particular epistemological stance of those involved in the creation, documentation, and organization of images of graffiti artworks.

References


A Formal Taxonomy of Knowledge Organization: Meta-Analysis and Facet Analysis

Richard P. Smiraglia*, Joshua Henry**, Elizabeth Milonas***, Chris Marchese****, Sergey Zherebchevsky*****

*Institute for Knowledge Organization and Structure, Inc., Lake Oswego, OR, USA
<richard.smiraglia@knoworg.org>

**Westminster Choir College, Princeton, NJ, USA <jhenry@rider.edu>

***Computer Systems Technology Department, New York City College of Technology (CUNY), New York, New York, USA <emilonas@citytech.cuny.edu>

****SUNY Nassau Community College, Garden City, New York, USA <marchesechris@yahoo.co>

*****Palmer School of Library and Information Science, Long Island University, Brookville, NY, USA <seniorpomodor@hotmail.com>

Richard P. Smiraglia is Senior Fellow and Executive Director of the Institute for Knowledge Organization and Structure, Inc. He is author of more than 300 publications in information science and knowledge organization. He holds a PhD from the University of Chicago and is Professor, iSchool at Milwaukee, Univ. of Wisconsin Milwaukee. He was 2018-2019 KNAW Visiting Professor at DANS, The Hague, The Netherlands, where he remains Visiting Fellow.

Joshua Henry received a dual master’s degree in library and information science and musicology in May 2017 from the University of Wisconsin-Milwaukee. He now works at Westminster Choir College as the Cataloging Librarian and is Associate Fellow of the Institute for Knowledge Organization and Structure, Inc. His research interests include film music and its interaction with concert music.

Elizabeth Milonas is an Assistant Professor in the Computer Systems Technology Department, New York City Technical College and Associate Fellow of the Institute for Knowledge Organization and Structure, Inc. She holds a BA in computer science and English literature from Fordham University, an MS in information systems from New York University, and a PhD from Long Island University. Her research interests focus on knowledge organization (KO), particularly in the result pages of web search engines. In addition, she is conducting research related to techniques for incorporating ethics in computer curriculum specifically in data science curriculum.

Christine Marchese, PhD is Associate Fellow and President of the Institute for Knowledge Organization and Structure, Inc Board of Directors; she is Associate Professor of business administration, SUNY Nassau Community College.

Sergey Zherebchevsky, MA, MSEd, MPhil in information studies, is Associate Fellow of the Institute for Knowledge Organization and Structure, Inc. He is a PhD candidate, Palmer School of Library and Information Science, Long Island University. His research interests include knowledge organization and knowledge organization systems and their integration in semantic web applications.
1.0 Introduction: domain-critical taxonomic underpinning

Taxonomy is the essential act of any science; at the most basic level, scientific activity is the identification, clustering, and disambiguation of empirical observations. The named list of any domain-specific set of observations is that domain’s taxonomy. In structured form, a taxonomy then is the authoritative representation of the core concepts in a science that serves as the basic reference set from which hypotheses may be developed and tested. It is in this manner that science moves forward by enriching its taxonomy through enhanced observations across time. Smiraglia (2014a, 51) wrote:

At the most basic level a taxonomy is an ordered list of terms together with their definitions or other determinant characteristics. Taxonomy is a way of defining the component entities in a domain ... The form and content of any taxonomy is dependent on the epistemology of the domain for which it has been developed.

and (54):

Rooted in empirical observation, taxonomies supply defining characteristics and identify the sources of the definitive science from which the characteristics were observed.

The science of knowledge organization as we practice it today was named by and evolved from promulgation by Ingetraut Dahlberg beginning in 1974 with the founding of the journal International Classification, followed auspiciously by the founding of the International Society for Knowledge Organization in 1989 (Dahlberg 2008, 82; Smiraglia 2014a, 40). Considered by its founder to be a science of science (Dahlberg 2006), in other words a meta-discipline, knowledge organization has constituted the formal application of “concept-theoretic” for the application of discovered concepts to ontological systems of various stripes. The activity of organizing knowledge has ancient roots, the mechanization and codification of which has occupied much of the period from 1839 (eg. Panizzi 1839; Jewett 1850; cf. Strout 1956) to the present (Smiraglia 2014a, 36-41). A hallmark of the maturation of a science is its production of formal systems for the representation of its theoretical positions. Such systems may take the form of ontologies or classifications, such as the 1993 “Classification of Knowledge Organization Literature” developed by Dahlberg (2006, 15). More recently, the online ISKO Encyclopedia of Knowledge Organization (IEKO) began to appear in 2017 (https://www.isko.org/cyclo/).

The science of knowledge organization was enhanced by the development of the methodological paradigm of domain analysis (Smiraglia 2015a), which involves the empirical identification of core concepts in every domain. The growth of knowledge of the domain of KO has been apparent from the increasing number of domain-analytical studies published about the domain (Smiraglia 2015b; 2015c). This effort, which involved overlapping studies of the literature and discourse of KO as a domain also embraced a fair amount of replication, thus generating the potential for further systematization of the core concepts in the domain of KO. In 2019, the Institute for Knowledge Organization and Structure, Inc. introduced the methodological meta-analytical tool of the domain analysis clinic or DAC (Smiraglia 2019), derived from

Abstract: Nearly fifty years after the incorporation of the International Society for Knowledge Organization and the introduction of its formal scientific journal Knowledge Organization, a comprehensive encyclopedia of the domain appeared. The practice of domain analysis for knowledge organization, twenty years after its introduction as a core methodology, has created the largest corpus of theoretical knowledge in the domain analysis of knowledge organization itself. A substantial body of research data, therefore, is available in the corpus of articles and conference papers reporting on the epistemological and ontological pillars of the science of knowledge organization. This paper is a report on the evolution of a formal taxonomy of knowledge organization, which is a product of an exhaustive meta-analysis of the KO domain. Our team compiled the corpus of twenty-nine formal published analyses together with key formative historical documents. We then analyzed the corpus thematically, bibliographically, and using co-word analysis to extract key concepts and the underlying faceted conceptual infrastructure. The taxonomy itself is faceted and is linked where possible to published definitions in the KO literature and as well as to the online ISKO Encyclopedia of Knowledge Organization. A dynamic project, the taxonomy will be maintained as linked open data and will grow as emergent research contributes new concepts or generates new facets.

Keywords: knowledge organization, domain analysis, terms, classification, taxonomy

Received: 1 June 2020; Revised: 17 August 2020; Accepted: 19 August 2020
Dahlberg’s (2006; 2011) idea of an institute devoted to the discovery of concepts but extended by the embrace of meta-analysis and empirical methods. From the first execution of the DAC came a formal taxonomy of knowledge organization that is the product of exhaustive meta-analysis of KO domain analytical research (Milonas 2019).

This paper reports on the evolution of a formal taxonomy of knowledge organization, which is a product of an exhaustive meta-analysis of the KO domain (Milonas 2019). Our team compiled the corpus of twenty-nine formal published analyses together with key formative historical documents. We then analyzed the corpus thematically, bibliographically, and using co-word analysis to extract key concepts and the underlying facets conceptual infrastructure. The taxonomy itself is faceted and is linked where possible to published definitions in the KO literature and as well as to the IEKO (Zherebchevsky 2019).

1.1 The meta-analytical corpus

Our attempt at an exhaustive meta-analysis had to start with a corpus that included analyses of the KO literature. Many studies have been conducted about the domain analysis of KO, but it was important to include papers that synthesized these in order to achieve an exhaustive meta-analysis. Three texts by Smiraglia (2013; 2015b; 2015c) provided the corpus with this synthesis. The breadth of his earliest work in the set is apparent when he explains (2013, 19-20) that:

In the 12 years or so in which KO has turned its attention specifically in a domain-analytical direction, there has been a modest increase in the number and frequency of studies specifically devoted to using Hjørland’s 11 approaches to the analysis of domains for the purpose of revealing their shared ontologies. In this chapter, we look briefly at those studies, which appear in KO’s three principal venues. Interestingly, we know from research (Smiraglia, 2011, 2012, 2013a) that those venues are primarily the biennial international conference proceedings from the International Society for Knowledge Organization (published in the series Advances in Knowledge Organization) and the journal Knowledge Organization, and a smattering of papers from the information science literature at large. Other sources are doctoral dissertations produced in schools of KO, although most of these are eventually reported formally in one of the other three venues.

His second study in the set synthesizes (2015b, 5):

17 studies of knowledge organization literature incorporated 3494 source papers, of which 1100 appeared in journals such as Cataloging & Classification Quarterly, Library Resources & Technical Services, or Library Quarterly, 444 appeared in Knowledge Organization, but 600 appeared in ISKO conference proceedings, and 1350 were papers in ISKO regional conference proceedings (the results of the meta-analysis are gathered in Smiraglia 2012b.)

Finally, his third study (2015c, 603) is “based on this core of 100 papers ... The 100 articles were contributed by 80 authors. Only 9 authors contributed more than one paper.” Although this last study might include the smallest amount of papers, Smiraglia points out (2015b, 610) that “this research is limited by the choice of which papers constitute the core of domain analysis, in KO, for KO. Other researchers might constitute the core differently. However, everything domain analytical in either the journal Knowledge Organization or the proceedings Advances in Knowledge Organization has been included.”

To further bolster the corpus with texts that contained core vocabulary that was likely to be considered taxonomic, we included Dahlberg’s seminal “Classification System for KO Literature” and the unpublished index to Smiraglia’s Elements of Knowledge Organization. These texts already provided a high-level view of KO, making them invaluable to our corpus.

Various domain analyses of ISKO or ISKO chapter publications also were included in the corpus. Although these papers are not syntheses of the KO literature, they provide a snapshot in time of the epistemological stance of KO from the primary organization for the scholarly discourse of KO. The authors of these papers used bibliometric and epistemological methods to analyze the KO domain internationally or regionally or to look at a particular aspect such as indexing (Guimarães and Tennis 2012) or the meaning of “concept” (Araújo, Tennis and Guimarães 2017). The various approaches share a common goal of understanding KO as a domain or stated differently (Castanha et al 2017, 8):

We understand that bibliometric studies are an important approach in domain analysis. Their use in combination with epistemological studies leads to better qualitative and quantitative analyses that take into account social and contextual aspects of indicators in order to better aid in the analysis. It also provides a valuable way to understand the information design, the visualization of the domain and the theoretical underpinnings of the social processes that permeate the information.

To round out our corpus, we included core writings of Joseph Tennis, former president of ISKO and the keynote speaker for the first DAC. These writings provided us with
a possible classification of KO research (2008), an ethical view of KO (2013), an understanding of classification theory (2005; 2015) and a way to approach domain analysis (2003). In regards to the last of these, Tennis suggests (193) two axes to conceptualize and delineate domains:

The Areas of Modulation, axis one, is an explicit statement of the name and extension of the domain examined. It states what is included, what is not included, and what the domain is called. Details as to how the domain is organized beneath this extension and name are the province of the second axis, Degrees of Specialization.

In total, our corpus constitutes a breadth of meta-level views of KO, which allowed us to conduct a meta-analysis of KO as a domain.

2.0 Corpus analysis for meta-analysis

To begin, meta-analysis of Dahlberg’s Classification System for Knowledge Organization Literature (CSKOL) and Smiraglia’s unpublished “Index” to the book Elements of Knowledge Organization was conducted. This meta-analysis was the first step in producing a formal knowledge organization taxonomy. Terms found in the two documents were analyzed, and unnecessary text (e.g., book section codes, introductory text, indefinite and definite articles, and conjunctions) was stripped. The remaining terms were uploaded into Voyant, an Internet-based open-source text mining tool (https://voyant-tools.org/?corpus=f02406bab97ca8671c8b9efb33cb1032b). The text mining tool identified the terms in order of their frequency within the stripped documents. Table 1 displays the most frequently occurring terms within each of the stripped documents. These terms include classification, indexing, systems, thesauri, knowledge, bibliographic, information, analysis, and social. The identification of these terms as the most frequently occurring is expected as these frequently occur in discourse related to the knowledge organization domain.

Terms with a frequency of five or higher were used as part of the facet analysis process. These terms were imported into an Excel spreadsheet. The terms from each of the stripped documents were placed in separate Excel columns. A side-by-side comparison of the terms in these two columns was conducted. Repeating terms were omitted and unique terms were identified and placed in a third column. Different word forms were regularized (e.g., “phenomena, phenomenon”). These common unique terms were examined as part of the facet analysis process.

Spiteri’s (1998) facet analysis model was then utilized. The facets were created with focus on the idea plane principles; differentiation (5), relevance (6), ascertainability (6), permanence (6), homogeneity (18), mutual exclusivity (18), and fundamental categories (18-9). In addition, when creating the facets, focus was also given to the verbal plane principles; context (11) and currency (11). As a result of the facet analysis, six facets and ten sub-facets were identified as follows; community (sub-facets: living things, society), tools (sub-facets: material, systems), action (sub-facets: methods, behavior, language), knowledge (sub-facets: concepts, subjects, of being), place, and time. Table 2 below displays the facets, sub-facets, and related terms.

An interesting yet predictable outcome from the meta-analysis and facet analysis of terms in both Smiraglia’s index and Dahlberg’s CSKOL is the absence of the significant and prominent themes of gender and identity. The project team, who were well-versed in both, was troubled by this apparent gap. A possible cause for this omission may be that although these documents present a fairly comprehensive view of the KO domain, they do not cover the breadth and width of the domain. Also, meta-analysis relying on frequency distribu-
<table>
<thead>
<tr>
<th>Facets</th>
<th>Sub-facets</th>
<th>Terms Smiraglia's Index</th>
<th>Terms Dahlberg CSKOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community</td>
<td>Living things</td>
<td>names, actors, author, public, plant, animal</td>
<td>human, persons, user, public</td>
</tr>
<tr>
<td></td>
<td>Society</td>
<td>society, communities, domain, culture</td>
<td>groups, societies</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Concepts</td>
<td>phenomena, phenomenon, concept, specificity, boundary, stance, perception, consciousness, cognitive, warrant, epistemic, epistemology, ontological, theory</td>
<td>religion, methodology, mathematical, standards, concepts, logic, policy, principles, kinds, concept, organizational, orientation, foundations, functions, level, standard, thought, trends, value, relations, shared, commodity, theoretical, theory</td>
</tr>
<tr>
<td></td>
<td>Subjects</td>
<td>science, subject</td>
<td>science, sciences, subject, fields, history, literature, mathematics, music, psychology, biology, sociology, ecology, economy, electronics, engineering, food, geography, geosciences, household, law, medicine, military, philosophy, physics, politics, sports, veterinary, biological, economic, mechanical, scientific, agriculture, arts, astronomy, chemistry</td>
</tr>
<tr>
<td>Of Being</td>
<td>physical</td>
<td></td>
<td>index, indexes, thesauri, data, information, work, computer, catalogues, programs, textbooks, books, bibliography, taxonomies, records, terms, codes, dictionaries, files, periodical, text, thesaurus, taxonomy, descriptors, manual, subjects, documents, glossaries , application, descriptor, editions, file, journals, materials, title, dictionary, documentation, headings, abstracts, archival, databases, document, internet, networks, forms, grammar, keywords, object, phrase, patents, pattern, record, specifications, storage, value, topics, word, content</td>
</tr>
<tr>
<td>Tools</td>
<td>Material</td>
<td>work, vocabularies, text, unit, term, resources, facet, element, document, descriptor, heading, content, code, categories, book, catalog, artifacts, apparatus, thesaurus, objects, network, metadata, map, entities, data, citation, catalogs, information, bibliography, sign, device , description, applications, taxonomies, ontologies, instantiation</td>
<td>systems, education, MARC, Bliss, Colon, Dewey Decimal System, Faceted Classification System, BBK</td>
</tr>
<tr>
<td></td>
<td>Systems</td>
<td>KOS</td>
<td></td>
</tr>
<tr>
<td>Action</td>
<td>Methods</td>
<td>FRBR, typology, taxonomy, documentation, ontology, notation, hierarchy, schema, rules, collocating, bibliographic, bibliographical</td>
<td>classification, documentation, indexing, method, bibliographic, taxonomy, technique, translation, procedures, reviews, guidelines, rules, hierarchy, ISBD, grammatical, standardization, copyright</td>
</tr>
<tr>
<td></td>
<td>Behavior</td>
<td>social, seeking, searching, retrieval, productivity, role, representation, intention, indexing, discovery, networking, discourse, classification, cataloging, analysis, activity, construction, controlled, control, descriptive, exploitative, consequences</td>
<td>social, indexing, organization, construction, evaluation, analysis, research, management, use, standardization, activities, mining, representation, coding, communication, labor, reclassification, relationships, retrieval, notation, teaching, translation, reviews, searches, abstracting, control, encoding, query, training, consistency, computerization, consolidation, concordance, linking, precision, search, selection, shelving, recurring, professional, authority</td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td>semantic, linguistic, language, semiotic, Symbolic</td>
<td>language, languages, terminology, semiotic, multilingual, semantics, syntax, linguistics, semantic, terminological</td>
</tr>
<tr>
<td>Places</td>
<td>Places</td>
<td>library, organization, universe, space, international, universal</td>
<td>universal, library, institution, online, international, countries, libraries, country, national, state</td>
</tr>
<tr>
<td>Time</td>
<td>Time</td>
<td></td>
<td>generation, current</td>
</tr>
</tbody>
</table>

*Table 2. Facet generation using Spiteri’s (1998) framework.*
tions of terms picks up the upper tier of a distribution but the granularity in the long tail does not get into the analysis. Alternatively, the lack of these important themes from these documents may reflect the recurring challenge of inclusiveness within KOSs (Olson and Schlegl 1999, Olson 2001, Fox 2016). The linear and hierarchical nature of these systems create an environment in which misrepresentation and elimination of themes are prevalent (Olson 2007). This pattern can be seen even in the widely used Dewey Decimal Classification where themes such as gender and identity are misrepresented or omitted (Olson 1998). Further study is warranted to determine the exact cause of the lack of gender and identity themes within the documents analyzed for this study.

3.0 Linking the Taxonomy of Knowledge Organization

A search for information within the corpus of the empirical research published in the KO literature inevitably turns up a spate of referential materials. Because we were looking for “definitions or other determinant characteristics” (Smiraglia 2014a, 51) of the taxonomic terms, the importance of achieving efficiency of retrieval seemed obvious. To accomplish this objective, a representative pool of KO literature—twenty-eight domain analytical articles and articles published in the journal Knowledge Organization (KO) from 1993 through 2019 were converted into digital text dossiers using Adobe Acrobat DC Pro v.9.1 computer software. The Elements of Knowledge Organization (EKO) (Smiraglia 2014a) was also available as a searchable e-book.

The initial search was done using the “find” and “full Acrobat search” functions of Adobe software within the merged dataset of KO literature. These automated search functions enabled the user to navigate from one instance of the term to another and view the list of all terms and its conceptual definitions within digital sets. However, the result of this first attempt to identify most relevant definitions was disappointing. First, although the search was conducted within a representative sample of the KO literature, some definitions describing basic KO terms were not found. Thus, a definition for the term “field,” which Taylor (1999, 242) defined as “A separately designated part of an encoded record; it may contain one or more subfields,” was not found in the corpus of KO journals. To overcome this problem, we expanded the pool of reviewed literature to include classic texts from Arlene Taylor, Carol Bean and Rebecca Green, Elaine Svenonius, and Richard Smiraglia.

Second, not all definitions available in the literature were contextually meaningful for our taxonomy. This finding might be explained by the fact that (https://www.isko.org/ko.html 2020):

Knowledge Organization publishes original research articles that: (1) clarify theoretical foundations (general ordering theory, philosophical foundations of knowledge and its artifacts, theoretical bases of classification, data analysis and reduction); (2) describe practical operations associated with indexing and classification, as well as applications of classification systems and thesauri, manual and machine indexing; (3) trace the history of knowledge organization; (4) discuss questions of education and training in classification; and (5) problems of terminology in general and with respect to special fields.

Therefore, this plethora of knowledge required careful examination to reveal out the most meaningful content. Third, important for developing a better searching strategy was the realization that most of the core terms could not be adequately defined with one basic conceptual definition. Once these considerations were taken into account, the searching strategy was modified. It was decided to search two types of definitions, conceptual and intentional, illustrating different aspects of the term usage. Both types were supposed to be meaningful in the context of our research field.

An example below shows definitions selected to describe terms “bibliographic/bibliography”:

Bibliographic classification – “a classification of knowledge and thought, and conversely, classification of knowledge [that] is available for a bibliographic classification” (Bliss 1952, 3). According to Bliss (1952, 2), such classifications are based on five principles: Subordination, Collocation, Maximal efficiency, Relativity of Classification, and Alternative location.

Bibliographic control – [is] “encompassing the creation, storage, manipulation, and retrieval of bibliographic data” (Smiraglia 1987, 2).

Bibliographic data – “information gathered in the process of creating bibliographic records” (Taylor 1999, 234)

Bibliographic entities – sets of individual documents that represent equivalence clustering of documents (Smiraglia 2014a, 13). “Sets have both abstract intellectual content and concrete semantic content” (Smiraglia 2014a, 26).

Bibliographic universe – “is a subset of all knowledge in which all instances of recorded and therefore potentially retrievable knowledge reside (Smiraglia 2001, 1).
Using modified search criteria described above, we obtained a sufficiently comprehensive and meaningful set of definitions that were used to populate our taxonomy.

4.0 Applied online taxonomy

The formal taxonomy is maintained online at (https://knoworg.org); a copy appears in the appendix below. Ongoing work includes continued linkage to the online IEKO, which is constantly growing. It also is our intention to create operable links to the Knowledge Organization System Observatory (KOSo) housed at the Dutch DANS (Data Archiving and Network Services) institute (Coen et al. 2019a-b). It is our intention to publish the taxonomy as linked open data (LOD) in the near future. A dynamic project, the taxonomy can then be maintained and continue to grow as emergent research contributes new concepts or generates new facets.

References


Smiraglia, Richard P. 2014b. “Index [to The Elements of Knowledge Organization].” Unpublished manuscript.
Appendix:

<table>
<thead>
<tr>
<th>Facets</th>
<th>Terms</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavior</td>
<td>Analysis</td>
<td>Denotes the domain analytical work within KO (Smiraglia and Lopez-Huertas 2015, 554) including domain analysis techniques, e.g., Citation Analysis, Co-word Analysis, Author Co-citation Analysis, Network Analysis, Cognitive Work Analysis (Smiraglia 2014a).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Domain analysis – (Hjørland 2017b); IEKO: <a href="https://www.isko.org/cyclo/domain_analysis">https://www.isko.org/cyclo/domain_analysis</a></td>
</tr>
<tr>
<td></td>
<td>Construction/structure</td>
<td>Construction – “...the construction of sets of terms (concepts) that are used in a specific community—or domain—mapped together with the relationships among them” (Smiraglia 2014a, 44).</td>
</tr>
<tr>
<td></td>
<td>Discourse</td>
<td>Structure – “a general framework or structure within which KOS can be built” (46).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discourse – the cultural action “by which language mediates knowledge” (Smiraglia 2014a, 27).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Epistemology is the division of philosophy that investigates the nature and origin of knowledge. In philosophy at large, epistemology is central because it embraces the theory of knowledge itself. ...The philosophical process engages a discourse in which skeptical challenges to any definition must be rebuked and therein lies the dilemma, for how can we study that which we cannot even define?” (20).</td>
</tr>
<tr>
<td>Facets</td>
<td>Terms</td>
<td>Notes</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Concepts/ideas/beliefs/theories</td>
<td>Documentation (see also Document)</td>
<td>“Documentation was a set of techniques developed to manage significant (or potentially significant) documents, meaning, in practice, printed texts” (Buckland 1997).</td>
</tr>
<tr>
<td></td>
<td>Boundary objects</td>
<td>“Terms used to pivot from one vocabulary to another” (Smiraglia 2014, 99).</td>
</tr>
<tr>
<td>Concept</td>
<td></td>
<td>Concepts are the building blocks of thoughts (Stanford Encyclopedia of Philosophy 2019).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“A concept is a knowledge unit” (Dahlberg 2010, 2946).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“A concept is regarded as the common element of both classification systems and thesauri” (Dahlberg 1974, 12).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“…knowledge is made up of concepts; … concepts can be ordered in diverse and useful ways (Smiraglia 2013, 2).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“The meaning (i.e., intension) of a term is the concept associated with that term” (Harney 2013, 135).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Concepts mean different things in different areas” (Hjørland 1997, 4).</td>
</tr>
<tr>
<td>Epistemology</td>
<td></td>
<td>“Epistemology is the division of philosophy that investigates the nature and origin of knowledge. In philosophy at large, epistemology is central because it embraces the theory of knowledge itself. The central problems for epistemology are the definition of knowledge, and the means of its acquisition” (Smiraglia 2014a, 20).</td>
</tr>
<tr>
<td>Phenomenon/Phenomena</td>
<td>Phenomenon (singular); phenomena (plural) – “A fact or situation that is observed to exist or happen, especially one whose cause or explanation is in question. (LEXICO Dictionary 2019)</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>“…the concept of specificity. The level of language to be employed is related to the intended functionality of the thesaurus. Specificity is related to the intended audience. …the more formal the language the more specific and precise the terms must be…” (Smiraglia 2014a, 81).</td>
</tr>
<tr>
<td>Theory</td>
<td></td>
<td>“Theory is a frequently-tested (and thereby affirmed) statement of the interacting requirements of a phenomenon” (Smiraglia 2014a, 7).</td>
</tr>
<tr>
<td>Language</td>
<td>Language</td>
<td>“Language – “A system which consists of a set of symbols (sentences) — realised phonetically by sounds — which are used in a regular order to convey a certain meaning. Apart from these formal characteristics, definitions of languages tend to highlight other aspects such as the fact that language is used regularly by humans and that it has a powerful social function.” (Small Dictionary of Linguistics)</td>
</tr>
<tr>
<td></td>
<td>Linguistics</td>
<td>Linguistics – “The study of language. (SmallDictionary of Linguistics)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semantic – “Relating to meaning in language or logic” (LEXICO Dictionary 2019).</td>
</tr>
</tbody>
</table>
| | | “Intension refers to the logical or definitional conditions that specify the set of all possible things a word or phrase could describe, while extension refers to the...
<table>
<thead>
<tr>
<th>Facets</th>
<th>Terms</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Semantic [extension, intension]</td>
<td>set of all actual things the word or phrase describes” (New World Encyclopedia 2018).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semiotic [sign] – “For Peirce, the sign consists of three components. These are the Representamen, the Interpretant, and the Object. The representamen is the concept as signal, the interpretant is the concept as reception, and the object is the concept as perception” (24).</td>
</tr>
<tr>
<td></td>
<td>Terminology</td>
<td>“The body of terms used with a particular technical application in a subject of study, profession, etc.” (LEXICO Dictionary 2019).</td>
</tr>
<tr>
<td>Material</td>
<td>Bibliography</td>
<td>Bibliography – “a list of the books referred to in a scholarly work, typically printed as an appendix” (LEXICO Dictionary 2019).</td>
</tr>
<tr>
<td></td>
<td>Catalog</td>
<td>Catalog – a complete list of items, typically one in alphabetical or other systematic order (LEXICO Dictionary 2019).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Catalog – “retrieval tool; an organized compilation of bibliographic metadata or an organized set of surrogate records that represent the holdings of a particular collection and/or resources to which access may be gained ...” (Joudrey and Taylor 2018, 625).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Library catalog – descriptive cataloging that applies a standardized set of rules, “currently RDA: Resource Description and Access, to record the title, authorship, and publication data for a work, describe the physical extent of the work, add bibliographic notes as necessary, and add access points for persons or entities associated with the creation of the work” (ALA 2019).</td>
</tr>
<tr>
<td>Categories</td>
<td>Citation</td>
<td>Category – “a grouping of people or things by type in any systematic arrangement” (Cambridge Dictionary 2019).</td>
</tr>
<tr>
<td>Citation</td>
<td>Citation</td>
<td>Citation – “A quotation of or explicit reference to a source for substantiation, as in a scholarly paper” (YOUR Dictionary 2019).</td>
</tr>
<tr>
<td></td>
<td>Citation Indexing – (Carina de Araújo, Gutierrez Castanha and Hjørland 2019).</td>
<td>IEKO: <a href="https://www.isko.org/cyclo/citation">https://www.isko.org/cyclo/citation</a></td>
</tr>
<tr>
<td>Document (see also Documentation)</td>
<td>Document</td>
<td>Document – “the physical container (an item) on which the text is recorded (Smiraglia 2001, 3).</td>
</tr>
<tr>
<td></td>
<td>Document – “an information-bearing message in recorded form” (Svenonius 2000, 8).</td>
<td>Document theory = (Buckland 2018)</td>
</tr>
<tr>
<td></td>
<td>Document</td>
<td>IEKO: <a href="https://www.isko.org/cyclo/document">https://www.isko.org/cyclo/document</a></td>
</tr>
<tr>
<td>Facets</td>
<td>Terms</td>
<td>Notes</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Information</td>
<td>Information is knowledge perceived. That is, what is contained in documents is potential information—it is recorded knowledge that may be consulted for whatever reason. But when that knowledge is consulted and is perceived by the human brain, information is the result. Information is a process and not a thing. Information therefore, is dynamic and not static. Information is what happens to a person when knowledge is perceived, because that new perception alters the person’s previously existing knowledge-base. (Smiraglia 2014, 11)</td>
<td>Information – “something received or obtained through informing” (Svenonius 2000, 7). “Metadata are descriptive terms that are applied to information resources, primarily for the purpose of facilitating retrieval” (Smiraglia, 2014a, 65).</td>
</tr>
<tr>
<td>Metadata</td>
<td>Metadata are descriptive terms that are applied to information resources, primarily for the purpose of facilitating retrieval” (Smiraglia, 2014a, 65).</td>
<td></td>
</tr>
<tr>
<td>Objects</td>
<td>… information objects, including not only books in libraries, but also representations of artifacts in museums and archival entities, as well as scientific models, ontological structures, and so forth. (Smiraglia 2008, 7).</td>
<td>Objects – “boundary objects,” or terms used to pivot from one vocabulary to another” (Smiraglia 2014a, 99).</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>“Taxonomy is a framework in which elements are defined, and categories are mutually exclusive and collectively exhaustive; …” (Smiraglia 2014a, 4). “… a taxonomy is an ordered list of terms together with their definitions or other determinant characteristics. … the form and content of any taxonomy is dependent on the epistemology of the domain for which it has been developed. In the generic sense, meaning the assignment of phenomena to specific categories, taxonomy is a form of classification. … taxonomy a highly specific sort of ontology, that arrives along with the definitions of the characteristics of the phenomena involved, and that also includes certain kinds of relationships, such as genus-species, etc.” (51).</td>
<td></td>
</tr>
<tr>
<td>Term</td>
<td>Term – “a generic term for a specific kind of recorded knowledge (Smiraglia 2014a, 70). Term – “although a word may have several senses, only one of them is intended when it is used as a term. Hence, a word is a term only when it designates one of its possible meanings” (Riggs 179, 152). “A term is a word or phrase used to denote a concept” (Pathak 2000, 29).</td>
<td>Terms – “In a specific discipline, items can be categorized and named as concepts of that discipline. The delimitation of knowledge into specific compartments is not easy, and as a result, in many fields of knowledge, specifically in the social sciences, where the same term is used in different discipline-specific contexts, the literature of that field provides the context in which a term is used and to which concept a term represents” (27) “In both computer science and information we see the construction of sets of terms (concepts) that are used in a specific community—or domain—mapped together with the relationships among them” (Smiraglia 2014a, 44).</td>
</tr>
<tr>
<td>Facets</td>
<td>Terms</td>
<td>Notes</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Textbooks [object]</td>
<td>Terms – “boundary objects, or points of opportunity for creating interoperable neighboring vocabularies from shared ontologies” (99).</td>
<td></td>
</tr>
<tr>
<td>Text</td>
<td>“An instantiation of a work” … whenever the work is manifest in physical form (in a book, for example).” (Smiraglia 2006).</td>
<td></td>
</tr>
<tr>
<td>Thesaurus</td>
<td>“A text is a set of semantic strings that communicate ideational content” (Smiraglia 2014a, 70).</td>
<td></td>
</tr>
<tr>
<td>Text</td>
<td>“… the set of words that constitute a writing. A text is not the same as a document, which is the physical container (an item) on which the text is recorded. A document may have only one text, but a text may appear on many documents.” Text, then, is another generic term that denotes the communicative aspect of the evidentiary value of a document.” (Smiraglia 2001, 3).</td>
<td></td>
</tr>
</tbody>
</table>
| Thesaurus | “A thesaurus is a complete knowledge organization system structured in natural language instead of according to its ontological construct. That is, the elements in a thesaurus are given in alphabetical order. Each term is then accompanied by a set of relational indicators that show its place in the overall hierarchy. Thesauri can be faceted, when terms from several facets are chosen and entered into a system in a string. Thesauri increasingly are multi-lingual to accommodate complex cultural demands” (Smiraglia 2014a, 79). | Thesaurus (for information retrieval) – Dextre Clarke 201

IEKO: https://www.isko.org/cyclo/thesaurus

“The term bibliography can have two definitions: there is bibliography itself, an activity, and there is a bibliography, the product of this activity. Bibliographies generally belong to two groups, one concerned with the listing of books and other documents, the other concerned primarily with the study of books as physical objects. ... It includes two specialities called systematic and enumerative bibliography ... The second group is concerned with the study of books as physical objects ... The several overlapping specialities in this side of the field include analytical bibliography, concerned with the ways in which specific books as physical objects were produced; textual bibliography, which uses these findings in the important work of establishing authenticity of content; and historical bibliography, which considers the relationships between a civilization and its books .... [The two groups] usefully come together ... most conspicuously in descriptive bibliography, concerned with the specification of particulars, based on the methods of analytical bibliography. (Krummel 1984, 4-5).

“It was around 1439 that Gutenberg created the mechanisms for printing from movable type that were to revolutionize the printing of books. We are looking, then, at the flowering of the marketplace for books only a bit more than a century after this remarkable invention. It was the need of the marketplace that drove the development of more sophisticated forms of bibliography.” (Smiraglia 2014a, 35)
### Facets | Terms | Notes
---|---|---
**Classification** | Classification – “the systematic ordering of knowledge” (Smiraglia 2014a, 48). | “By the middle of the twentieth century Clapp (1950) was writing that bibliography was one of the arts of communication found at a second level of utterance, treating prior records of communication, and in need of patterns of effective arrangement.... In the same volume, Jesse Shera and Margaret Egan referred to social role of bibliography as part of the problem of inter and between group communication (1950, 17) (Smiraglia 2014a, 40).”
| Classification – “the placing of subjects into categories; in organization of information, classification is the process of determining where an information resource fits into a given hierarchy and then assigning the notation associated with the appropriate level of the hierarchy to the information resource and to its metadata” (Joudrey and Taylor 2018, 626). | Classification – (Hjørland 2017a)
| IEKO: https://www.isko.org/cyclo/classification | “Classifications of characteristics of phenomena, and these need not be mutually exclusive nor collectively exhaustive” (Smiraglia 2014a, 53).
| “The term typology is used for the same sort of arrangement when the entities involved are called types instead of taxa. Typologies are used in anthropology, archaeology, linguistics, theology, and psychology. In most instances, typologies are less robust scientifically than taxonomies, which means a type is assigned based on empirical observation but always is subject to change given analysis from future observations” (53). | Classification – (Hjørland 2017a)
| IEKO: https://www.isko.org/cyclo/classification | Indexing – “intellectual analysis of the subject matter of a document (2.15) to identify the concepts (2.11) represented in it, and allocation of the corresponding index terms (2.26) to allow the information to be retrieved” (ISO 2011, 5)
| The process of creating surrogate records, especially the access points for information packages; such work done in commercial enterprises is often called indexing, while similar work done in not-for-profit agencies is usually called cataloging (Taylor 1999, 244). | Indexing: Concepts and theory – (Hjørland 2018)
| IEKO: https://www.isko.org/cyclo/indexing | Method – “a systematic procedure, technique, or mode of inquiry employed by or proper to a particular discipline or art.”
<p>| <em>Merriam-Webster Dictionary. 2019</em> |</p>
<table>
<thead>
<tr>
<th>Facets</th>
<th>Terms</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology</td>
<td>Ontology – “a domain of thought in philosophy. In philosophy ontology is the study of being—of what is ... ontology allows us to isolate certain principles of physical vs. metaphysical, of categories and the entities that are their contents, of the relationships among all of the above, of attributes of phenomena such as facts, properties, energy, space, time, etc.” (Smiraglia 2014a, 43).</td>
<td></td>
</tr>
<tr>
<td>Of Being</td>
<td>Knowledge – “that which is known” (Smiraglia 2014a, 3).</td>
<td>Class 92 covers selected items of knowledge organization literature. It is found in the cumulative database of International Society for Knowledge Organization (ISKO 2019). Principal actors in the domain – knowledge producers and knowledge users (Smiraglia 2014a, 16).</td>
</tr>
<tr>
<td>People/living things</td>
<td>Persons and institutions in KO</td>
<td>In the context of domain analysis – a producer of knowledge (Smiraglia, 2014a, 16). The primary metric for measuring the scientific productivity of an author in domain analysis techniques, e.g. Citation Analysis, Author Co-citation Analysis. Whatever we consider to be the most basic element of reality, we deem to be things or, more formally, entities” (Bean and Green 2001, 3).</td>
</tr>
<tr>
<td></td>
<td>Actors</td>
<td>Entity – “a term used in the field of knowledge organization to indicate an item; both “entity” and “item” are used in order to avoid using “book” or other such specific designation” (Taylor 1999, 242).</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>“A bibliographic entity is a unique instance of recorded knowledge (e.g., a dissertation, a novel, a symphony, etc.)” (Smiraglia 2001, 2).</td>
</tr>
<tr>
<td>Society</td>
<td>Domain (see also Analysis)</td>
<td>“A domain is a group that shares an ontology, undertakes common research or work, and also engages in discourse or communication, formally or informally” (Smiraglia 2014a, 85). “A domain is best understood as a unit of analysis for the construction of a KOS” (86). In the context of knowledge organization, the “social” refers to “the confluence of art, commerce, and technology... [that] come together at important moments to act as a collective catalyst to move the domain forward (Smiraglia 2014a, 33). In KO, the prevailing point of view is that “the growth of knowledge over the whole course of human history” (34) and the way knowledge is organized is shaped by the social realities of the world.</td>
</tr>
<tr>
<td></td>
<td>Social</td>
<td>Disciplines – (Hammarfeldt 2019)</td>
</tr>
<tr>
<td></td>
<td>Subject</td>
<td>Subject – one of “the attributes of a given bibliographic condition... such as “origin” or “subject” the better to define the intension of each set over against the intensions of the other sets” (Smiraglia 2014a, 13).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subject (of document) (Hjørland 2017c)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IEKO: <a href="https://www.isko.org/cyclo/discipline">https://www.isko.org/cyclo/discipline</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>IEKO: <a href="https://www.isko.org/cyclo/subject">https://www.isko.org/cyclo/subject</a></td>
</tr>
<tr>
<td>Facets</td>
<td>Terms</td>
<td>Notes</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Systems</td>
<td>Universal classification</td>
<td>Universal classification – “one that applies the same approach and terminology across domains” (Szostak 2014, 161).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Universal classification – “should be considered the sum of a number of domain specific systems (birds, cars, countries, religions, sciences, etc.).” (Hjørland 2017 447).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Universal classification – “…bibliographic classifications such as the Dewey Decimal Classification or the Universal Decimal Classification” (Smiraglia 2014a, 52).</td>
</tr>
</tbody>
</table>

**Taxonomic References**


Carina de Araujo, Paula, Renata Cristina Gutierrez Castanha and Birger Hjørland. 2019. “Citation Indexing.” ISKO Encyclopedia of Knowledge Organization https://www.isko.org/cyclo/citation


Hjørland, Birger. 2017c. “Classification.” *Knowledge Organization* 44: 97-128


Hjørland, Birger. 2017c. “Subject (of Documents).” *Knowledge Organization* 44: 55-64


Towards a Model of Urban Studies Classification

B. Preedip Balaji* and M. Dhanamjaya**

*Indian Institute for Human Settlements Library, 197/36, 2nd Main, Sadashivanagar, Bangalore—560080, Karnataka, India, <pbalaji@iihs.ac.in>
**Reva University, Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bangalore—560064, Karnataka, India, <registrar@reva.edu.in>

B. Preedip Balaji is Consultant at the Library at Indian Institute for Human Settlements, Bangalore and is a doctoral student in the School of Library and Information Science, Reva University, Bangalore. He has a master’s degree in library and information science from Bishop Heber College, Bharathidasan University, Trichy. His research interests are open access, domain analysis, digital libraries, semantic web and knowledge organization, and he can be reached at pbalaji@iihs.ac.in

M. Dhanamjaya is Pro Vice-Chancellor at Reva University, Bangalore. He has a PhD in library and information science and management having research interests in scientometrics, information management, open access, public libraries and scholarly communication. He is also a guide for library and information science at Reva University.


Abstract: Evolution of cities is a subject of research for over a hundred years in the organization of urban knowledge systems. Locating five key methodological approaches used by urban scholars and practitioners, this paper demonstrates different relationships between urban studies and classification. Five significant themes form the background of urban studies literature. The first theme sources and literature explore organizing urban materials into sources and literature with a unique dimension of spatiality. The second theme discusses three important facets: scale as a geographic unit of analysis and space as an abstract entity and system as a set of interdependent parts of urban places. The third theme, known as “other” urban, argued for the poor treatment of global south and how it builds inclusivity. The fourth theme, classification and retrieval, investigates the relationship between urban materials and user needs. The last theme, classification schemes, highlights subject treatment of urban in the existing library classification schemes. This paper concludes that the five themes discussed point to a model of urban studies classification. However, this model is not just concerned with urban methods, facets and formats, but explores how each theme interconnects with various sets of people—urbanists, practitioners and librarians—and through studying these actors, established boundaries of urban theories, urban librarianship and knowledge organization are crossed.

Received: 18 May 2020; Revised: 30 August 2020, 13 October 2020; Accepted: 19 October 2020

Keywords: urban studies, classification, sources, literature

1.0 Introduction

The classification in urban research has a long history in urban studies, especially for urbanists, researchers, practitioners and librarians who are trying to understand its theoretical and practical aspects of knowledge organization. The basic subjectivities of urban studies at their foundations remain conceptually elusive and evolving, since cities are delimited in their physical, economic and social terms. Lacking easily identifiable boundaries, the definition of cities is contested and varies across countries and geographies, while globally meaningful classification schema for urban studies is indispensable for knowledge management in urban librarianship (Scott and Storper 2015; McGranahan and Satterthwaite 2014). This significance necessitates understanding what the characteristics of urbanism are and how to organize it as a production and evaluation of special classification as Hjørland (2019) suggested through domain analysis and as a paradigm in domain knowledge organization systems (Smiraglia 2015, 3-4).

Despite efforts of understanding urban knowledge, urban studies remain as an interdisciplinary domain having no canonical base and consensus around its ontological core, but has built epistemological diversity and its necessary tol-
erance to it as Paddison (2015) discussed. Urban studies as a field of study has its complexity, heterogeneity and interconnections, where the influence of planning, social theories and geography dominated their epistemologies in the past century. However, the proliferating growth of city planning and a corpus of global diversity of urban literature deploying methodical, disciplinary and sectoral views necessitated the organization of it as an engaging area of research on what constitutes urban studies (Brenner 2009). This influx to the urban studies demands understanding the challenges of urban classification and to look at new ways of offering views into the future as Pissourios and Lagopoulos (2017) pointed out for its applications and use cases.

In understanding urban studies as a typical classification scheme of study, this study analyses subject treatment of urban concepts and facets of urban research but cannot be comprehensive. The purpose is to highlight historical features and description, followed by a discussion on issues and processes for improvement. This approach considers arranging real-life objects in practice not just for subject access, hence examines arranging urban knowledge in a specific library and will be used for real-life libraries for use among librarians, urbanists and practitioners. This is also equally relevant to online classification, for instance creating shelf-listings, subject guides and subject headings.

This paper begins with an outline of five different types of methodology distinct in urban studies literature, where each methodology signifies the relationship between classification and urban studies. Next, five of the main themes in urban studies: sources and literature, facets, “other” urban, classification and retrieval and classification schemes are discussed. Through these methodologies and themes, we demonstrate their interrelationships between urban studies and classification to explain an emerging model of urban studies classification.

2.0 Methods

Analysis of urban studies literature reveals that there are five demonstrated methodological approaches: sector, discipline, method, geography and program (IIHS 2019). These five methods in urban classification draw from urban theories and classification to unpack the concepts, key terminologies and their relationships in urban context used by authors and practitioners in urban research. Sector refers to a sociological, economic or political subdivision of a society specific through an economic area or activity in urban areas and in “spatial planning requires continuous horizontal integration across sectors and vertical integration across scales to support the development of integrated cities and territories” (United Nations 2017a).

Discipline in urban studies have overarching and inherently developed areas of research conducted at multi-, inter-, transdisciplinary and converging levels from sciences, arts to social sciences (Advisory Committee for Environmental Research and Education 2018). For example, the disciplines that dominate the urban studies are design of the built environment (e.g., architecture), policy (e.g., public administration) and social science (e.g., sociology). In a milieu of disparate urbanism, method as an instrument is to investigate cities and city life from multidisciplinary clusters, and, therefore, methods in urban research are built on various schools of thought (Classical School, Chicago School, Frankfurt School and Los Angeles School); theories of human ecology, neo-Marxism, ethnographic methods, spatial analysis, urban history and on contemporary debates around postcolonial, planetary and provincialized urban theories. Urban data and tools (computing, programs and retrieval), medium (audio-video), scale (a conceptual arrangement of space) and form (physical characteristics) are the other methods critical for urban classification.

Being conceptual and action-oriented, the other two methodologies are used less. Geography uses study of urban spaces, urban ways of being and how to approach city and taking part in the intellectual and political stance of critical urban geography. As part of the urbanized world, it means highlighting and participating in attempts to change cities for the better (Jonas et al. 2015). Since it is not defined by one paradigm or canon of work, it further connotes social, cultural, human, political, economic and labor geographies. Program refers to a set of events and initiatives that steer the urban policies at the national, subnational and global level, significantly shaping the history of urban policies. For example, the Vancouver Declaration on Human Settlements (Habitat I) set the stage for urban settlements development and future programs such as Habitat II and Habitat III (Habitat 1976). Most significantly, further on a global scale, programs such as Millennium Development Goals (2000-2015), Sustainable Development Goals (2016-2030), New Urban Agenda (2016-) push for sustainable urban development and are aimed to ensure that cities are inclusive, safe, resilient and sustainable communities (United Nations 2017b).

3.0 Major themes of urban studies

In our analysis of how urban studies classification has evolved, the following five important themes have been found as central in urban studies literature—forming the main body of the paper. This section adapted the five themes framework Lee (2012) proposed in domain modeling as below:

3.1 Sources and literature

Approaching urban study materials into sources and literature is fundamental in urban studies classification, which
looks at classifying them within urban research based on their form. As common among library classification schemes, Shapackov (1992) stated that all subject matter takes form and Dahlberg (2008) positioned form in International Coding Classification as “general form concept,” which is a basis to categorize subject areas. In urban studies, major sources are government documents, images and media, cartographic materials and GIS and statistics and data. Since subject representation is captured through form as a common subdivision, human settlements and their sources of site, data, area and period specifically look at the characteristics of cities, metropolitan regions and urban and suburban areas, whereas the urban literature covers a vast diversity of sources and genres, for example urban literary genres. Since the broad sources of urban documents are fiscal (budget and financial reports), architecture and planning documents at the city, local and regional levels, it is essential to make this distinction as there is a need to have these two prime categories to organize all items, where sources are primary documents as published works and literature is used to mean works whose subject is urban.

This distinction of sources and literature leads to discuss other important ideas about urban classification. Dividing sources and literature can have practical issues, as against conceptual arrangement, since arranging them by format might be preferred by users. As Lee (2012) cited that though multiple other formats exist to integrate into library classification, this is, however, dependent upon the sources and literature divide: practical versus conceptual and medium versus format. Understanding this deepens the representation of this divide, including various types of sources, beyond the two-dimensional materials (e.g., geometric shapes) to other three-dimensional objects like artists’ books (arts collections) or globes (cartographic collections), which also constitute as sources.

The diversity of sources and urban literature makes it challenging, when there are no pre-existing library schemes to understand, sources and literature divide in urban librarianship. An exception to urban studies is the JEL Classification Code (AEA 2020), which includes collective works and volumes, subject handbooks, and all the other unclassifiable objects to organize scholarly literature in economics incorporating urban aspects within the JEL Code (see their classification in Section 3.5).

Once the sources and literature decision is made, there are two different ways to make this division. First, addressing the debate of sources and literature for settlements and their classification. In urban studies discourse, human settlements formed in site and situation have features that are its population size, density, occupational structures, administrative boundaries and functional activities. Second, using attributes such as pattern, size and housing density and if combined with spatiality, then they are categorized as urban and rural settlements—which is fundamental to organization of urban studies classification. Since the definitions of settlement geography have been rather inadequate, UNSD (2017) cites how challenging are the concepts of locality, urban and rural areas breakdown in urban classification. This is with reference to arranging population and housing census documents and in reporting vital statistics systems, given the distinct population clusters and national differences respectively. Some of the core concepts in urban studies are demonstrated using Dahlberg’s categories in domain analysis (2009, 172) in Table 1.

### 3.2 Facets

Different aspects of sources and literature are used for classification, but urban studies literature focuses on select key facets more. Although medium and form are the common facets—scale, space and system—emerge as important facets for arranging sources and literature. This is specifically for investigating the levels of data in a hierarchy (from metropolitan to local); conceptual arrangement of a location as an abstract-concrete idea from space to place (from India to Bangalore) and how different parts of a whole system function together (built infrastructure, communities and natural environment). Again, this also concerns with understanding scenarios when existing tools of measurement are getting reworked for governing the cities. Being centric around how space is absolute for the physical materials of the world and socially produced, this basically refers to a set of relationships between society and the spaces that society produces. Since cities pose challenges understanding their

<table>
<thead>
<tr>
<th>Category</th>
<th>Concept</th>
<th>Attribute</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity</td>
<td>Settlements</td>
<td>Part_of</td>
<td>Abstract-concrete</td>
</tr>
<tr>
<td>Property</td>
<td>Urbanism</td>
<td>Approach_of</td>
<td>Characteristics</td>
</tr>
<tr>
<td>Activity</td>
<td>Urbanization</td>
<td>Process_of</td>
<td>Cause-effect</td>
</tr>
<tr>
<td>Dimension</td>
<td>Urban system</td>
<td>Function_of</td>
<td>Whole-part</td>
</tr>
</tbody>
</table>

*Table 1. Categorical relationship in urban studies.*
properties and relationships between concepts, urban practitioners, scholars and librarians face difficulties in organizing urban categorization, reclassification and their cartographies amidst scales of growth as to how urban as a space has evolved and why using system as a way of thinking to achieve multiple benefits in outcomes became inevitable.

Scale is a conceptual arrangement of space and is thought in terms of levels—local, national, global—as a key element to understand space. As an appropriate unit of analysis, scale is used by geographers to indicate that geographical processes operating at different levels (e.g., global, national, regional, neighborhood and household). It is increasingly recognized that scale as a process is interconnected in complex ways such as drawing lines on maps and allowing certain activities to take place within certain territories (Jonas et al. 2015, 319). According to Bowen and Gleeson (2019, 116) central place theory, time and space are signifiers for human settlements from hamlets to megacities, since theorizing premodern, modern and postmodern cities in time and cities in space involves economic, political and social conditions on a local, regional and global scales. Library users approach cartographic materials and urban data by their scale and hence arranging these sources by scale would be ideal in urban classification. For example, to arrange geo-data, theme and place name, datatypes are used as facets.

Space as a conceptual arrangement brings to light urban as a space of production having inherent inequalities, of places/flows, public and private use and commodified. Referring to a territory in abstract and concrete terms as an idea, space is a term often used in a general sense to indicate geography, location or distance but also used specifically by urban geographers to acknowledge the socially constructed nature of environments (e.g., gendered space, public space and green space). With advancements, this increasingly connotes the social and digital spaces as one of the main divisions of urban studies involving physical and cyberspaces and the technological innovations in these spaces, examples are urban design and smart cities. In urban classification, understanding user needs of space is essential, since spatiality as a facet determines the level of classification from global, national to local sources of urban documents as well as how users approach physical and digital spaces in global cities (NIUA 1988; Mainka et al. 2013).

The United Nations (2017a) puts system as a whole-part of interdependent elements of urban as an entity to achieve multiple benefits in outcomes at multi-scales as a system of system to reinforce the primacy of the relationship between elements and the flow of materials and energy rather than individual elements. As the Advisory Committee for Environmental Research and Education (2018) defined urban systems science captures the spatial and temporal variations in the character of urban nodes or settlements—in any area and the interactions between them, intersecting with built infrastructure, natural environment and communities. Coffey (1998) describes this as composition of a “set of elements (cities) and the interactions (social, economic, financial, informational and so on) within or between them,” having influenced and contributed to urban systems research, which in turn used as hard and soft domains of research (Neirotti 2014) in urban classification and significantly contribute to modelling different areas of specialization. For example, urban energy systems and urban education.

### 3.3 “Other” urban

Since the dominance of urban theories are from the global north, subject treatment of urban from the developing countries as Lawhon (2020) states create space for the inclusion of southern cities in urban theory from the global south, which has increasingly gained attention in urban classification. This has been particularly concerning to authors decolonising the western social theories with the treatment of southern urbanism. There are several factors for this paradigm shift and why the classification of “other” urban is problematic.

The evolving landscape of the global south within urban studies and how classification schemes should be able to keep up with these rapid changes is a possibility. In locating the essence of urbanity, Schindler (2017) viewed the southern urbanism as a distinctive type of settlement, differing from its global north counterparts. Kong and Qian (2017) argue about the Anglophone dominance of urban studies, and that knowledge production from developing countries is dynamic and contentious in urban discourses, issues that were less known but have been collectively transforming globally. Hence, urban classification schemes need to be equally dynamic to cover this epistemic new urbanism.

Beyond the western urban theory, urban literature approaches southern theory for its applications and perspectives notwithstanding Anglo-American hegemony. Issues with the classification of global south theory is not just with the materials of classification schemes but their inherent structure. This is partly because the traditional classification schemes were written before the global south became mainstreamed within urban studies. Therefore, library schemes should be designed to coherently present views from the global south to be accommodating, inclusive and as representative as possible. The possibilities and limitations of this is felt within urban studies, as the history of urban has structured gaps and biases with north-south inequalities and library classification schemes are not designed to reflect these effects. Examining this, Connell (2014) stated how the southern theories are not a fixed set of propositions but have strong imperatives for inclusivity. Further, Lawhon and Truelove (2020) critiqued the southern urban
theories of geography towards broadening pathways and possibilities for "more spatially diverse and theoretically robust urban studies (e.g., northern/southern, Euro-America/postcolony and global urban studies)."

 Differential treatment of non-western urban works will have consequential effects if the context and relationships of the subject are not understood in urban classification. As an example, from developing countries for persons working at the landfills in waste management, can we categorize them as rag pickers or refuse collectors? It is important to use appropriate terminologies and subject headings from a social justice point of view to understand this as much as from a classification view to categorize people (e.g., urban poor) in urban classification. Here applicability, culture and context of concepts comes to the fore, which may not embody or exist in developed countries like these additional instances such as urban informality, migrant labor and slums. Parker (2015, 35-36) noted how Booth’s categories of urban poor in four classes is one classic example based on distinct characteristics of certain topographical areas in the city of London, characterised by living conditions. Another prominent example is urban educator versus urban practitioner, where a significant quantity of urban literature looks at the urban education as a reflective practice but has less discussions as a practitioner face serious criticism. It is very important to discuss these challenges, which only flexible and adaptable approaches of classification can deliver in knowledge organization.

3.4 Classification and retrieval

According to Martin et al. (2003), the relationship between the materiality of research and the language we use to represent urban knowledge is multiple, having parallels across different disciplines and hence cannot have a unitary meaning of a given place, neighbourhood or city. For knowledge discovery of urban materials, accurate retrieval sits at the core of the classification, where it should be expressive, flexible and extensible to be a retrieval-based classification. Since urban studies is one of the highly interdisciplinary subjects, the approach of users is difficult to perceive and to interpret the queries of various users for processing and representation in retrieval should be savvy. Some examples are cities and climate change, gender and planning, cities and towns in literature, urban economics and urban engineering, which warrant assessing the requirements of both subject analysis and keeping the interests of users in classification.

NIUA (1988, v) in organizing a bibliography noted that urban studies were organized combining disciplines and sectors and then into subject and themes. This is explained further for users this way: at the first level of classification by theme and getting in-depth at the second level by subthemes and at the third level by a spatial dimension, for example, from national to state level documents. Broadly supportive of a critical urban studies, urban literature concentrates upon classification of medium and form of urban sources, which are not limited to images or datasets for retrieval. Urban classification requirements include spatial resolution requirements, urban land cover/use, spatial images, urban sounds, urban scene and applications of GIS and urban informatics to map urban areas and data processing of urban sources. Hence, the classification and retrieval of urban materials need to organize in the following key areas:

1. Natural resources: use case scenarios and models such as land cover and use, water, energy, air and biodiversity.
2. Physical attributes of built environment: landscapes, types, form, structure, pattern, size, proportion, morphology and scene.

Classification and retrieval should be able to accommodate the localization, since urban studies has many local sources and require local action at the municipal level. For example, local climate change action planning requires the inclusion of retrieval at the regional and state level since climate action cannot be too restrictive at the local level in urban governance. Authority control of persons, non-governmental, bilateral and multilateral organizations should be considered, where academia, practice-based, consultancy-led projects and global organizations are key stakeholders in urban research and practice. Additionally, practitioner, urban scholar, urbanist and policy maker do not have the same level of associative relationships in the urban research, but have different needs; urban practitioner and urbanist were not used much in the earlier generations but are mainstreaming now.

Urban studies have many materials, especially in literary genres where the “subject less” is common and have variety of genres difficult to understand in classification, for example, green humor and urban tales. Nonetheless, this extends as a subject beyond the topical subject headings (e.g., Urbanization—India), subject as a region (e.g., Latin America—Developing countries), subject as era (e.g., Urban policy—21st century) and subject as a genre (e.g., sources, designs and plans, exhibitions and case studies) and should be assessed for its implications in retrieval of urban sources.

3.5 Classification schemes

As Lee (2012) pointed out, there are three classification systems followed in domain KOSs: a section of general scheme, a special scheme for the subject or a home-grown scheme
designed at an individual library. Delimitation of the urban places and their classification into physical, historical, sociocultural and functional classes determine the following three urban classification schemes prevalent in literature: census classification, subject classification and city classification. These classification schemes in the urban literature are categorized further as below:

1. Classification in use case scenarios (Pissourios and Lagopoulos 2017). Urban/rural classification; urban use case scenarios for land use, zoning, urban settlements and structure types; land use/cover planning ontology and urban ontologies.

2. Urban innovation and engineering in digital and physical spaces of global cities (Mainka et al. 2013) and core infrastructures such as energy, water, solid waste, buildings and transport for city planning (Neirotti 2014).

3. Application of technology for cities and applied urban environments for sustainable development and smart growth (e.g., sustainability indicators, smart city rankings, urban resilience frameworks) (National Research Council 2010).

From the quantity of these classification schemes what widely accepted is that no classification can fit for all geographies as Pateman (2011) argued. First the conceptual, disciplinary and thematic boundaries are highly distributed; organization of urban systems are complex as Bretagnolle et al. (2009) cited, and deciphering the characteristics of urban/rural areas are challenged for their definitional, economic, political and governance underpinnings, while the existing generic library classification schemes for urban studies are rather generic. Settlements in library classification schemes are organized under different disciplines. For example, settlements are classified under geography in UDC (911.37) and LC classification schemes (GF 101-127); in DDC, it goes under social sciences (307.76) and the arts (711); and UDC treats urban in the arts (71) disciplines (Dewey et al. 2011; UDC Consortium 2013).

With disciplinary diversity, specialty and heterogeneity, Balaji (2019) found that the subject treatment of urban in library classification schemes is fragmented. Among all the classification schemes, by far the richest classification covering urban concepts is the DDC. Urban regions under Table 2 Areas (— 17 Socioeconomic regions) by concentration of population (— 173), classify urban areas into urban regions, suburban regions and rural regions as three classes. This notation specifies urban areas that can be added to main class numbers in DDC23, standardizing their socioeconomic status and/or combining any main subject with urban properties (Dewey et al. 2011). UDC (UDC Consortium 2013) under the common auxiliaries of place—Table 1e—categorize urban areas as political and administrative units. According to Mills and Broughton (1977), in the Bliss Bibliographic Classification auxiliary schedule 2, places classify urban as regions by land and resource use and population.

As a subject, the context of urban is treated differently in library classification schemes. In DDC and LCC, communities were strongly represented as part of the development and planning in social organization, whereas UDC has a geographical focus of settlements. JEL Classification Code followed by the American Economic Association (AEA 2020) has an urban focus on following areas of economics as a subject classification (see Table 2).

### 4.0 Conclusion

Urban studies domain in all its complexity and richness is one of the contested landscapes of cultural representation and social experience globally, helping to model urban studies classification. The first theme, sources and literature, discusses various sources of urban materials in classification and how spatiality is a third dimension unique to urban classification. Second, facets in urban studies classification are scale, space and system, which explore how these three facets provide a useful exploration of analysis of urban areas in terms of hierarchy, abstract ideas and elements. “Other” urban includes the literature from the global south, which
is for mainstreaming but also has strong imperative in urban discourses for inclusion. Classification and retrieval show the different needs of users for retrieval of urban sources, highlighting the relationship between users and needs. The fifth theme elaborates on how the pre-existing schemes of library classification deal with urban studies as a subject. As demonstrated above, what emerges from this analysis is an interesting set of core concepts drawing on five methodologies and themes in urban studies classification. The five themes discussed point to a model of urban studies classification, based upon the ontological and epistemological relationships, intersecting with the disciplinary boundaries of urban theories, urban librarianship and knowledge organization. Our further research will identify key hierarchical themes and research areas within the urban studies domain for a specialist classification.

References


An applied ontology-Oriented Case Study to Distinguish Public and Private Institutions Through Their Documents†

Jaime A. Pinto* and Maurício B. Almeida**

*Federal University of Minas Gerais (UFMG), School of information science, Av. Antônio Carlos, 6627, Campus Pampulha, ZIP 31.270-901, Belo Horizonte, Minas Gerais, Brazil, <jaimeandradepinto@gmail.com>

**Federal University of Minas Gerais (UFMG), School of information science, Av. Antônio Carlos, 6627, Campus Pampulha, ZIP 31.270-901, Belo Horizonte, Minas Gerais, Brazil, <mba@eci.ufmg.br>

Abstract: The institutions we create shape many of the activities we engage insofar as they are pervasive entities in our society. In an era full of new technologies, including the semantic web, there is a movement toward sound conceptual modeling for socio-technical solutions applied to government institutions. To develop these complex solutions, one needs to deepen the ontological status of entities in the institutional domain, because literature is full of ambiguous and ad-hoc hypotheses about distinctions between public and private corporations. We believe we can find better explanations for such distinctions in the interdisciplinary field of library information science. Within an ongoing semantic web project, we focus on a study case of official documents. First, we analyze theories about public and private corporations, seeking a reliable ontological distinction between them; then, by focusing on documents produced by each type of corporation, we hope to provide a well-founded analysis. Second, we adopt the aforementioned theories and the new analysis as recommendations for the improvement for the access and understanding of public documents, through appropriate classification of them within government information systems. This project, ultimately, aims to maximize the transparency of public government documents by favoring retrieval and comprehension by a society with plenty of automated information systems.
1.0 Introduction

Institutions have made significant investments in information and knowledge management initiatives, mainly through the development of information systems with increased sophistication. In developing these complex systems, in addition to several technological resources, one needs to study the ontological status of the entities that populate the corporate domain.

In this work, the operational definition of a domain will be applied to the electronic government (eGov) as a concept (Smiraglia 2016). This definition is derived from several empirical works and included the following main points: i) an ontological base with an underlying teleology; ii) a set of common hypotheses; iii) an epistemological consensus; and, iv) a reasonable consensual semantics. Within an ongoing semantic web project, which aims to increase the transparency of public government documents, we take advantage of the interdisciplinary aspect of library and information science (LIS) to develop a theoretical perspective through a case study of government budget documents. Our ultimate goal is to improve the access, retrieval, and understanding of key government documents by the society as a whole.

To reach this goal, our starting points, then, are studies in the field of applied ontology (Smith 2008, 2014; Smith et al. 2007). First, we analyze interdisciplinary perspectives on public government institutions and private business corporations. One might say that there is no difference between public and private corporations, but some believe that they are fundamentally different because each has a distinct purpose.

Thus, we first seek this ontological distinction on which to build a hypothesis focusing our analysis on documents. Documents are a core subject of LIS, and we believe that the documents produced and used by institutions are one of their most distinctive and illuminating features. Second, we apply the framework thus acquired to a semantic web project for electronic government (eGov) systems.

Key terms used here are “institution,” “corporation,” and “ontology.” In this paper, we follow Berle and Means (1932) in using the term “corporation” to denote modern institutions originating in the US, public or private. The word “ontology” as used here is a combination of metaphysical principles and computational techniques used in applied ontology (Munn e Smith 2013). We also use the phrases “ontology as discipline” and “ontology as artifact” to distinguish the two meanings of the term (Almeida 2013). As a complementary view, “social ontology” is the term adopted to name studies within the ontology of social entities (Tuomela 2013).

To reach our goals, the paper is organized as follows. In Section 2, we introduce speech acts theory and document acts theory; in Section 3, we present the historical nature of institutions from interdisciplinary literature and our analysis for comparing private and public institutions; in Section 4, we describe a case study that focuses on documents for creating a model for government institutions; and finally, in Section 5, we discuss our findings and offer final remarks and suggestions for further research as well.

2.0 Background of social ontology

Since the 1990s, the use of the term “ontology” within computer science and information science is used to name both models as well as the theoretical principles that underlie those models. While the “spatio-temporal ontology” corresponds to an attempt to classify natural categories (Munn e Smith 2013), the “social ontology” is a kind of theory for social entities, which addresses human artifacts and social devices (Searle 2010). Sections 2.1 and 2.2 address ontology as a social theory, while Section 2.3 presents it as an artifact.

2.1 Speech acts: what can one do with words?

The theory of “speech acts” was proposed by John Langshaw Austin (1911-1960), a British philosopher of language, for whom one can “do things with words” (Austin 1962). For Austin, these forms of speech acts, which he considered the basic units of meaning, are constituted by three connected dimensions: locutionary acts, illocutionary acts, and perlocutionary acts. For example, in the proposition “I promise to pay you tomorrow,” there is an utterance (the verb promise) that constitutes the act of promising rather than the description of mental states. When a person utters the sentence, the promise is concretized; in other words, the force that characterizes the act is the promise (Almeida, Silva, and Brochhausen 2017).

The importance of speech acts in institutional contexts relies on their ability to bring about new social entities as obligations and claims to which promises and orders give rise. In Searle’s theory, institutions are created by a simple function: X counts as Y in C, where X is an entity and Y is the status of such entity in a context C (Searle 2010). For example, Trump (X) counts as president (Y) in the United States (C). Institutions are a particular case of this formula in which an issue emerges; there is no member X as a physical entity. This issue...
is solved by a divide called a “standing declaration;” they are of a certain kind that allows one to make something happen simply by promising that it will really happen. In the case of institutions, executing and fulfilling the statute of incorporation X counts as the creation and maintenance of a corporation Y within a specific legal code C (Searle 1976).

Thus, the institutional reality is created through speech acts by which duties and obligations are delivered throughout an institution, unfolding new forms of social interactions. However, the evanescence of speech acts, a consequence of their inherent orality, limits their possible impacts on institutional environments (Smith 2012).

2.2 Document acts: what can one do with documents?

While speech acts exist only in the moment of their performance, documents can also convey acts, allowing them to persist in time even while absorbing modifications through the document’s life cycle. This is the premise of the theory of “document acts” (Smith 2012); promises and obligations are established through speech acts.

The theory of document acts describes how people and institutions use documents to bring about new social entities. A document, as an input of a document-act, is the bearer of social and institutional powers (ethical and legal) that cause a variety of social effects. A document-act is a kind of process, for example, a contract (i.e., a document) brings about an obligation (e.g., to pay some amount of money) through a document-act (e.g., the process of signing the document); a statute of a corporation (i.e., a document) brings about to existence a corporation (e.g., with the obligation of delivering specific product or service) through a document-act (e.g., the process in which government stamps the statute of a corporation).

The modern institution requires a kind of document that endures over time, working as an input of acts that fulfill the obligations in different situations. We address here those additional functions of documents that, rather than only record information, creates claims and obligations for people, as well as new social facts in society.

2.3 The Ontology of Document-Acts

The document acts theory was implemented for practical purposes through “document act ontology” (hereafter “D-Acts Ontology) (Brochhausen, Almeida, e Slaughter 2013) using pre-existing resources of the Basic Formal Ontology (BFO) (Smith et al. 2007) and the Information Artifact Ontology (IAO) (Ceusters 2012). While IAO provides alternatives for representing information artifacts as documents that record information, D-Acts Ontology represents the kind of document used as an input of a document act to deliver obligations.

Figure 1 presents an essential branch of D-Acts Ontology that concentrates on the roles required to trigger document acts. Figure 2 presents another branch of D-Acts Ontology that gathers documents and the acts they contain. These two branches are linked by the class, “Declaration,” that appears in both Figures 1 and 2. Classes depicted by a shaded rectangle are not D-Acts Ontology classes, but BFO or IAO.
classes that illustrate the connection between the middle and the top-level ontologies.

D-Acts Ontology assumes that claims and obligations are subtypes of BFO’s classes. Within this complex network of roles, acts, and realizations, one can explicate the basic operations of institutions. Using documents, people can trigger new processes while registering them for future assessments of responsibilities and performance.

3.0 Corporations: public or private?

In this section, we focus on finding the distinctions between two kinds of institutions, public and private corporations. The roots of institutions are briefly explained in Section 3.1 and extended through excerpts of interdisciplinary literature in Section 3.2. In Section 3.3, we present our formulation for a distinction, using the speech and document acts theories.

3.1. The nature of corporations—a brief overview

An understanding of institutions as legal entities had already existed when the English crown began to charter business organizations in the fifteenth century (Williston 1888). In this context, institutions have certain core attributes, which were adopted by jurists in America, where an institution began to possess additional legal attributes (Jones 1994). The classical formulation of institutional attributes...
has come to be known as the “artificial person” doctrine, one of the several that arose to explain the notion of the corporate personality (Iwai 2007, 2010, 1999). US law maintains that a private or a public corporation must be treated as a person. Debates in countries like France, Germany, and Italy led to the emergence of different theories (Machen 1911): fiction theory, concession theory, group personality theory, bracket theory, purpose theory, Hohfeld’s theory, and Kelsen’s theory.

3.2 Interdisciplinary perspectives on private vs. public corporations

From the perspective of economic and management sciences, fields inherently interested in institutions, one sees significant contrasts in stakeholders and sources of funding. By definition, public organizations are funded by the citizens, while private organizations are funded by owners and shareholders. Indeed, issues regarding funding have a considerable impact on governance practices. The management field addresses the public/private distinction but does not present any fundamental characteristic that differentiates the two. Corporations operating in the private sector have many differences relative to public sector government units and agencies. Financial management, on the other hand, identifies as the source of the most significant differences: funding sources and the character of stakeholders (Allison 1992; Ciepley 2013; Meier e O’toole 2011; Rainey e Chun 2009).

Within the literature from information systems, one can find a variety of initiatives on international cooperation and research addressing corporate information systems in the context of semantic web. Studies like this do not present any theoretical concerns with public/private distinctions, even though identified as “ontological” (European 2016). Indeed, the use of conceptual vocabularies such as “public” and “private” often generates much more confusion than understanding, because each suffers from theoretical premises with contextual, historical, and temporal assumptions and connotations (Weintraub 1997).

Much of the contemporary debate is still characterized by the ambiguity of post-modernism biases, which prevents the clear development of alternatives (Ackroyd and Fleetwood 2000). There are even lines of thought that reject a clear boundary between private and public institutions (Ciepley 2013, p.139).

In addition to fields related to management, social ontology offers alternative methods for explicating distinctions between public and private institutions. The analysis of private organizations include two main dimensions: i) the descriptive dimension, which describes how to divide corporations into units and subunits; and, ii) the prescriptive dimension, which explains duties, obligations, and responsibilities that corporations have to manage (Brochhausen, Almeida, e Slaughter 2013). Theories introduced so far (Sections 2.1 and 2.2) can be used to formulate new alternatives.

3.3 Private and public corporations: are they different?

Ad-hoc views, including some presented here in previous sections, are not suitable for ontological studies that take a more serious and precise approach to the nature of entities as a key criterion for constructing a classification system. To advance an ontological analysis, here we try to understand the identity of entities.

The criterion of identity is usually attributed to the German philosopher Friedrich Ludwig Gottlob Frege (1848—1925), who asked how one can know whether “a” is identical to “b,” when “a” and “b” are entities. To answer this question from an ontological perspective, one should refer to the properties that objects of the same category share, to the extent that they are identical. From this perspective, the distinctive property is described in terms of essences—an Aristotelian principle embedded in the computational artifacts we adopt here, namely BFO, IAO, and D-Acts Ontology. In seeking essential properties, we rely on documents and their function-in bringing institutions into existence. Before presenting a final analysis, we need to present four preliminary distinctions.

The first distinction held is between natural things and human artifacts; for example, an orange is a “natural thing” and a car is a “human artifact.” Human artifacts are not natural things, since they maintain a trace of human intentions that were applied to their design, like cars, hammers, and software (Preston 2013).

The second distinction involves the word “public,” which has here the sense of something that can be known to all individuals in society. Artifacts depend not only on one human being or his individual intentional state but on the intentionality that transcends their creator (Thomasson 2009). For example, if someone creates a spoon, that spoon is recognized as something used to eat food. One can create your own “spoon” by folding a piece of cardboard, but even though it can function as a spoon, it would not be recognized as a “real” spoon: an object created for a specific purpose, by adult people, in a particular context. Thus, an institution is a human artifact describable by two aforementioned dimensions (Section 3): 1) the descriptive dimension, which is a spatio-temporal perspective; and, 2) normative dimension, which is a social perspective. The normative dimension is relevant here, because documents are artifacts working as inputs for document-acts, which convey rules and norms, through which people create and maintain institutions.

The third distinction, between individual and collective intentionality, develops the need for public norms. Collective intentionality is more than a collection of individual inten-
tional states; this fact explains much of the cohesion of a society (Tuomela 2013). In short, the creation of an artifact is subject to certain norms, because artifacts are recognizable as something that has to be used (i.e., applied) in a certain way (rather than in other ways) by an intended community and in a collective context (Dipert 1993). Considering the scope of human artifacts created for collective purposes, we follow Thomasson (2009) in considering the existence of public artifacts as the result of public norms in a process highly dependent on cultural aspects, despite the intentionality involved.

The fourth and last distinction addresses regulative versus constitutive rules (Searle 2010), which we employ to better explain the notion of a “public norm.” The former rule, which merely regulates, reflects acts performed independently of the rule. Constitutive rules not only regulate behaviors but also create the possibility of the existence of the same behavior they regulate. For example, “driving on the right side of the road” is a regulative rule that limits driving behaviors in which the act of driving is separated from the rule. On the other hand, there is no possibility of the game outside of chess rules; that is, the rules of chess create the possibility of chess.

Searle (2010) uses this distinction to explain how institutions are created by institutional facts that only exist through collective intentionality in the scope of a complex network of constitutive rules. For example, “John is a driver licensed to drive in the US” is an institutional fact that can exist only within a constitutive rule system collectively accepted within the US. One relevant last consideration of Searle’s framework is the kind of constitutive rule called a “standing declaration,” a declaration that has a doubled direction of realization. Searle (2010) explained the double direction within speech acts, and Smith (2012) extends it to documents explaining that a document, in one direction, is able to make a thing comes to existence; a thing, on the contrary direction, can make changes in documents. For example, following the direction of realization human-mind to the world, a blueprint brings a building to existence; the building (follow the opposite direction, from human to mind) make changes in the original blueprint by replacing it with an “as-built” blueprint. This framework allows us to elaborate comparisons between public and private corporations, which constitutes the remainder of this section.

First, institutions, public or private, are human artifacts as they depend on human intentionality for their creation. As human artifacts, they must be public in the sense that they should be recognized as something associated with social use. To the same extent that a spoon could be made of cardboard (i.e., defined purely by function), one can also create a private corporation and maintain it for only personal use. In contrast, a public corporation cannot exist as a secret, and by definition, it cannot be a private tool.

Second, standing declarations make something happen by representing the possibility of the very same thing happening. For example, the existence of someone who satisfied the condition of being the oldest son of a dead king, in the context of medieval Europe, brought a new king to existence (Searle, 2010). By the same token, standing declarations also explain the creation of institutions: one declares that anyone who makes a declaration of a certain kind (in the context of a complex system of rules) will have constituted a corporation. Institutions, however, cannot be treated in the same way as the case of kings in old Europe, because they require explicit rules in the context of a complex legal structure and, thus, are dependent on written language. Documents have the power to accomplish these requirements. Such documents work as inputs for document acts that can create institutions.

Finally, the kind of document-act that creates a public corporation is different from the kind that creates a private corporation. While the former is a law, the latter is a statute of the corporation. Then, if we should point one distinctive and unique property of a document that makes evident the difference between public and private corporations, we choose the deontic powers embodied in the documents that create and maintain an institution. Besides, while a deontic power that creates public corporations emerges from society as a whole, the deontic power that creates private corporation emerges from a small group or only one individual.

4.0 Case study: a government budget

In this section, we present an application of the theories described thus far to illustrate their practical utility. Section 4.1 describes how data was collected as part of our experiment; Section 4.2 explains the experiment itself.

4.1. Collecting data: choosing a government document to test

To explore the characteristics of documents in corporations, we chose a budget. One can make budgets for several purposes, but “federal budgets” are public documents that work as a tool for planning by containing both revenues and expenditures within a given period. According to Brazilian law, the “Brazilian Federal Budget” (BFB) should be publicly available through an open linked data format for purposes of transparency. Thus, this public document made available is not the paper document; instead, a semantic model called the “Brazilian Federal Budget Ontology” (BFBO) (Ararú, Santos, and Silva 2015).

The website “Ontological Model for Expenditure Classification of the Brazilian Federal Budget” (available in Portuguese at http://vocab.c.gov.br/2013/09/loa) provides documentation as well as all BFBO datasets since 2000. We
collected data about the predicted expenditures from this website.

4.2 The experiment: building a new model

According to the ontological artifacts presented before (Section 2.3), we try to match of BFO’s top-level classes to BFBO classes. BFBO reveals a poor classification structure. It is, in fact, just a list. It does not provide any organization in the form of generic classes but starts with the root concept “thing” and only uses representation languages typical from semantic web initiatives. Also, classes in BFBO are not adequately defined; for example, “year” is a property and not a class (Figure 3). In an even more simplified view, it can be said that the BFBO could have more semantic expressiveness if it were constructed as a thesaurus in which the relationships of the related term type (RT) were enriched. In the way that thesaurus is today, only the relations of the narrow term type (NT) and the broader term (BT) are supported. Studies show that the enrichment of RT relationships can help to understand and to retrieve information (Wu 2010).

To perform our experiment, we made BFBO classes fall under D-Acts and IAO models while maintaining BFO as the top-level. To follow the BFO framework and the kinds of concepts involved, some explanations of D-Acts and IAO models are necessary:

– An “information content entity” is an “object” that is about something, for example, as the content of a book can migrate from a paper book to a digital file.
– One level below this, a “directive information entity” is an “information content entity,” for example, a consent letter provides a nurse the right to extract blood.
– One more level below, a “document” is an “information content entity,” characterized as a collection of information content entities intended to be understood together.

On the other hand, IAO’s “directive information entity” also encompasses devices that realize processes. Thus, a “directive information entity” is a kind of document that realizes acts, which is exactly the sense of documents within the D-Acts ontology. Under IAO’s “directive information entity,” we found a class “plan specification” that we have the sense of “planning,” characteristic of a budget. In the context of D-Acts, a plan specification works as an input for a document act, which, in this context, represents all instances of Brazilian law involved in budget enforcement and approval.

From this, we created a class called “public budget,” a subclass of plan specification as the starting point to better organize BFBO (Figure 4). With these and other amendments to the original BFBO, a new structure reflects a more suitable proposal of classification. By “suitable,” we mean that the classes of the computational artifact received theoretical considerations to be classified and not defined in an ad-hoc way. Finally, it is worthwhile to say that we do not expect that formal frameworks do justice to the richness of the word “document” in the context of LIS.

4.3 Testing the new model

One of the most relevant developments in seeking a formal framework that can be applied in eGov systems is the attempt to solve problems of interoperability between systems. As this work is part of an ongoing research project, the evaluation of the proposed ontology will be performed at another stage of the project. The planned assessment is a data integration experiment using BFBO as an intermediate ontology in the context of ontology-based data access (OBDA) methodology. Another possibility of validation is the interoperability between KOS, using the already existing vocabulary and taxonomies of the Brazilian government. Several research paths have proven promising to facilitate the implementation of ontology (Lei Zeng e Mai Chan 2004). The remodeling experiment presented in the previous section is part of an ongoing project in which we have constructed a complete practical interoperability case. In...
such an on-going project we: i) modify the public budget document using the ontological frameworks provided, adding new classes required to better represent official documents; and, ii) demonstrate the use of the new structure to retrieve data from BFB originally stored as a relational database.

5.0 Summary and conclusions

This paper introduces an ontological view that enables one to understand better, distinguish, and design eGov solutions. We began with an examination of the ontological status of institutions, emphasizing the weaknesses of ad-hoc hypotheses about the distinction of public-private. We focused our analysis on documents and used a formal framework to gather together theories of applied ontology and to present examples of applications.

We concluded that the kind of document-act that creates a public corporation is different from the type that creates a private corporation; while the former is a law, the latter is a statute of a corporation. Thus, this is evidence, a qualified one, of the differences between public and private corporations. We also highlighted that deontic powers embodied in the documents are responsible for creating and maintaining institutions.

The practical case verifies the possibility of using a public-domain ontology to build eGov solutions. We introduced D-Acts Ontology, which implements the theory of document acts and allows its practical use to describe how people and institutions use documents to bring about new social entities; we used a corresponding ontology to define classes within the socio-technical universe.

Finally, we hope to have provided a sort of template of tasks for designers and system developers for all their offices and public bodies, containing a solid theoretical foundation for the study of the eGov projects through ontological frameworks. By using this framework and its associated tools, designers can take new perspectives and challenge heterogeneity issues within existing and new systems. In future research, we hope to provide an account of our experience with legacy systems in the context of the semantic web, in preserving such systems, and offering better services to citizens.

References


Gercina Ângela de Lima*, Maria Luiza de Almeida Campos**, Patrícia Lopes Ferreira França***

*Universidade Federal de Minas Gerais (UFMG), Programa de Pós-Graduação Gestão & Organização do Conhecimento (PPG-GOC), Av. Antônio Carlos, 6.627, Belo Horizonte, MG, 31.270-901, Brazil, <glima@eci.ufmg.br>
**Universidade Federal Fluminense (UFF), Programa de Pós-Graduação em Ciência da Informação (PPGCI), Rua Tiradentes, 148, Ingá, Niterói, RJ, 24210-510, Brazil, <maria.almeida@pq.cnpq.br>
***Serviço Federal de Processamento de Dados, Av. José Candido da Silveira, 1200, Belo Horizonte, MG, 31035-572, Brazil, <patricialopesvirtual@gmail.com>

Gercina Ângela de Lima is a full professor at the Information Science School of Federal University of Minas Gerais, Brazil. She holds a PhD in information science from UFMG (Brazil) and a master’s degree in library science and services from Clark Atlanta University (USA). She is the coordinator of the Research Group MHTX Research Group—Conceptual Modeling for Organization and Representation of Hypertextual Information and has a research scholarship from The Brazilian Research Council (CNPq) since 2010. Her current research interests include organization and representation of knowledge, hypertext, digital library, and web semantic.

Maria Luiza de Almeida Campos holds a PhD in information science and is Researcher and Professor at the post-graduate program, at Universidade Federal Fluminense and Universidade Federal da Bahia, Brazil. She is the coordinator of the Research Group EOOCI—Ontic and Ontological Studies in Informational Contexts and has a research scholarship from The Brazilian Research Council (CNPq) since 2005. Her areas of interest are knowledge organization, models and theories of knowledge representation, terminology, semantic interoperability, and foundational ontologies.

Patrícia Lopes Ferreira França is a PhD student at Federal University of Minas Gerais where she also holds a master’s degree in information science. She worked as Pedagogical Supervisor and Instructional Designer at Senac Minas. She is a member of the MHTX Research Group—Conceptual Modeling for Organization and Representation of Hypertextual Information.

Abstract: This article presents a set of principles for knowledge modeling in knowledge organization systems in specific domains. It discusses the representational problem, comparing the abstraction mechanisms present in the theories related to representation in concept systems, taken from foundational authors of information science, computer science, and terminology approaches. Parallel to this context, several representational possibilities arise to assist the modeler in the activity of elaborating models of representation. It describes the application of theoretical and
methodological principles when organizing, representing, and managing navigation on learning paths in the corporate education field. As a concept proof, it exposes a conceptual model of learning paths and discusses a literature review on this subject to verify to what extent these principles are being applied. It concludes that we can consider the principles discussed in this study as relevant, since they expand the modelers’ freedom, not making him hostage to a specific model.

Received: 16 April 2020; Revised: 6 August 2020, 2 September 2020, 9 October 2020, 13 October, 2020; Accepted: 19 October 2020

Keywords: knowledge representation, knowledge organization, learning paths, principles

† The authors wish to thank the National Development Council Scientific and Technological (CNPq / Brazil) for the granting of the Research Productivity-PQ.

1.0 Introduction

This study proposes a set of principles for the elaboration of domains’ conceptual models, based on the theories of information science, computer science, and terminology. Such principles allow the representation of knowledge in different domains to find a common core of essential principles to model knowledge. In this approach, domain, according to Hjørland (2002; 2004), makes up a discipline, a community, an application field, or even an action within a community.

The conceptual domain models are present in several knowledge organization systems (KOS) and builds a systematic structure. In this sense, this investigation discusses the representational problem, comparing the abstraction mechanisms present in theories related to the representation of concepts’ systems, from authors linked to the areas of information science, computer science, and terminology.

In information science, Ranganatha’s faceted classification theory (1951; 1967) and the concept theory formulated by Dahlberg (1978a; 1978b) stand out; both allow the representation of knowledge domains. In computer science, authors investigate the representational models associated with the modeling of fundamental ontologies (Guarino 1998a and 1998b; Smith 2004). Based on Terminology, established principles are used to determine concepts and their relations presented by Wuester (1981). With the foundations proposed by these authors, in the three areas of knowledge, this study lists principles that aim to assist the modeler in his activity of elaborating models of representation.

In recent years there were quite a few systematic studies of the theories underlying the areas related to the construction of conceptual models. These models allow the development of documentary languages, computer systems, hyper-texts, and systems aimed at building knowledge bases also called specialized systems and ontologies (Campos 2001; Lima 2004; Lopes 2018). Thus, it should add several representational possibilities to think about a given reality, from a theoretical-methodological stance that allows the modeler to overcome specific models of representation and reflect on the principles underlying the modeling process.

In this study, we understand modeling from the basis developed by Le Moigne (1977) with the theory of the general system also called modeling theory. Le Moigne affirms that to know is to model; the knowledge process is equivalent to the construction of the world/domain models that will be built, making it possible to describe and provide explanations about the observed phenomena. The author, when proposing the development of a methodological stance, asks the researcher not to think about the diversity of models but mainly about the principles that make modeling possible.

Thus, the information professional must also be aware of the principles adopted when determining the modus operandi for the elaboration of a KOS. This is because, according to Le Moigne (1977), in any modeling theory used in the elaboration of representation models, the model, product of this representation, cannot be a definitive and immutable truth but must result from a given process, of a particular construction, or of a worldview linked to space and time. Therefore, knowledge is the action of its building, a process in which the principles adopted for this construction must be clear.

Therefore, aiming at the validation of the established principles, namely the reasoning method used to analyze the domain, the representation object, the relationships between objects for the representation of conceptual structures, and the forms of graphic representation, these principles were applied in corporate education. In this domain, the focus was on the organization and representation of learning paths based on two procedures: 1) the first investigated, in this subject literature, to what extent such principles were usable; and, 2) the second was the development of a conceptual model of a learning path for corporate education as a concept proof.

We can understand learning paths as a systematic and multimodal set of learning units, containing different navigation schemes. These schemes can vary from linear and prescriptive models, passing through more hierarchical models and reaching network models in which navigation is
freer (Lopes and Lima 2019). In corporate education, these paths tend to contribute to better use of the resources invested in lifelong learning, besides making it possible to overcome the teaching approach called “one size fits all” (Adesina and Molloy 2011; Freitas and Brandão 2006; Lubchak et al. 2012; Suazo et al. 2012; Yang 2012).

This article consists of four sections, including this introductory part. In Section 2, with the title “Knowledge representation and principles for domain modeling,” the principles regarding domain modeling is discussed based on foundational authors taken from the concept of knowledge representation. In Section 3, entitled “The representation of learning paths in the context of corporate education,” we present the definition of learning paths and to what extent the literature on procedures for the development of learning paths uses the principles of modeling explored. In addition, as a concept proof, the third section applies these principles in the development of a conceptual model. Finally, in Section 4, we present our conclusions based on the results achieved throughout the study.

2.0 Knowledge representation and principles for domain modeling

Several scientific fields define the term knowledge representation differently. A well-known and updated definition by Sowa (2000), conceived under a theoretical approach (philosophical and logical), considers knowledge representation as the application of logic and ontology in the tasks of building computer models for some domains. Artificial intelligence research discusses knowledge modeling or knowledge engineering methods (Studer et al. 1998). We obtain another point of view through linguistic studies and their elucidation in knowledge representation (Löbner 2002). In cognitive science, knowledge representation is based on the individual’s world model, the context and the way one sees the world individually, comprising the sum of different knowledge structures (Gardner 1996; Mey 1982). From the perspective of information science, the authors Cleveland and Cleveland (2001), Lancaster (2003), and Stock and Stock (2013) treat knowledge representation as an approach to solving problems, such as structuring, storing information, and how to find and recover them accurately and effectively. It is important to note that among these definitions presented in several domains of science, Sowa’s (2000) view is restricted to the analysis proposed in this article.

Knowledge representation is a recurring theme in several sciences, looking into how the processing of knowledge occurs in the human mind and how to materialize this knowledge. The symbolic form of this representation is something that concerns the scientific scope of documentation since its origin. Such representation associates with ways of expressing information, as highlighted by Vickery (1978). The problem is relevant and extends to many other situations besides documents and indexes. It is necessary to decide how to represent knowledge so that these representations can be manipulated (Vickery 1986, 145) in the structure of software data, the syntactic and semantic structures of natural language, the representation of knowledge in artificial intelligence, and the models of human memory.

In this sense, San Segundo (2004, 109) highlights that if knowledge is an integration process, knowledge representation will be “concepts, theories, models, formats, descriptions and structures that have a meaning of the information symbolization and, more recently, electronic information.” Knowledge organization is a necessary process, which becomes more urgent as the volume of information increases. The major aim of this process is to organize the knowledge of a domain and make it available for later retrieval. The organization and retrieval of information face the same challenge to meet the needs of potential users.

In Lima’s view (2020), to organize knowledge in any area, from representation to recovery, we first need to study the concepts that make up this knowledge field and the relationships between them. According to Lima (2020), we can modify the concept based on prior knowledge on the topic after a mental elaboration, transforming from the information unit to a conceptual unit for communication purposes. Thus, knowledge representation is a process that uses terminological instruments as products. So, it is necessary to establish guiding principles in the construction of these systems as described in the following section.

2.1 Principles for modeling knowledge domains in knowledge organization systems

Conceptual modeling consists of the creation of models of representation/description from a worldview (conceptualization) that exists over a domain of knowledge. This practice focuses on “identifying, analyzing, and describing the essential concepts and restrictions of a domain” (Guizzardi et al. 2002, 65). For Campos (2004), the modeling process thus requires the displacement of the “phenomenal world” to a space where knowledge representation mechanisms allow formalization processes of objects and their relationships in pre-defined representation contexts.

In order to provide the modeler or classificationist with a range of representational tools for modeling domains and, based on a systematic investigation of methods and theories, we arrived at the determination of five fundamental principles for the act of modeling domains of knowledge, which are present in studies developed within the scope of representing learning paths in the corporate education context. The first principle concerns the “domain of knowledge to be modeled.” The second concerns the “reasoning method” used for the knowledge organization within a domain. The
third one analyzes the definition of “the object of representation,” identifying the knowledge unit to be represented. The fourth concerns the “relationships between objects,” aiming to verify the possibilities of semantic connection/separation between the concepts of a given domain. And the fifth shows the “forms of graphic representation” that a model can adopt. In this section, we will present each of these principles, establishing a dialogue between the fields of information science, computer science, and terminology.

2.1.1 Knowledge domain to be modeled

The first principle to consider when developing more appropriate methodologies for conceptual models should be the determination of the knowledge domain. Therefore, it is important to check if the subject covers disciplines that have a paradigmatic or syntagmatic cut in the meeting of knowledge units. It is necessary to verify if there is evidence of the matter in an already established canonical area or if it is part of an area that gathers knowledge from different canonical areas.

When you insert the domain of the treated subject in a paradigmatic cut, it presents the units of knowledge, in most cases, hierarchically. If it is the syntagmatic type domain (having over one subject area of a canonical cut), its units of knowledge will form several hierarchical trees that will come together from generic planning. It is important to emphasize that the development of a conceptual model is complex, which requires knowledge about the subject and about the community that will benefit from the modeling. Thus, it is important that methodology developers for conceptual models establish a stage that can provide discussions about the type of subject thematic section treated in the conceptual model that is being created and the target audience of the intended mapped domain.

2.1.2 Reasoning method

The second principle concerns the adoption of a reasoning method used for the knowledge organization within a domain. Every knowledge organization, according to Morin (2000, 24), follows principles and rules in which organization can include connection operations (conjunction, inclusion, and implication) and separation (differentiation, opposition, selection, and exclusion). The process is circular and cyclical, carried out in stages, moving from separation to connection, from analysis to synthesis, and from synthesis to analysis. Therefore, knowledge involves simultaneously, separation and connection, analysis and synthesis.

Although there is an understanding that the organization of knowledge must occur from a systemic perspective, it is possible to identify the development of distinct organizational models. Traditionally, in computer science, models were developed using two different methods: 1) the inductive method, also called bottom-up; and, 2) the deductive method, also called top-down.

The inductive method allows the elaboration of models that begin from the attributes’ context. Therefore, it starts by representing the elements/objects and the relationships of a context, as to obtain an organized structure of concepts, which is grouped according to the classes they belong to, according to their attributes. The deductive method, on the other hand, proposes the elaboration of abstraction mechanisms in order to first think about the domain/context regardless of the consideration of elements and their relations (this would be a later step). In this sense, a process of knowledge logical division follows, starting from general categories, containing generic concepts to more specific concepts. In effect, it generates a set of hierarchical classes in which each subclass of elements is a set of the immediately preceding class. We can associate these methods with theories defended in different knowledge areas.

2.1.3 Representation object

The third principle focuses on the represented object. In the context of the analyzed theories (general terminology theory, formal ontology theory, faceted classification theory, and concept theory), the object is defined as the smallest unit of manipulation/representation in each context. In each of these theories and in the methodologies used, it is possible to identify what can be considered an object of representation and the importance of that object as the smallest representation unit in addition to the implications that result from this definition.

Regarding terminology theory, Wuester defines this object of representation as a concept, which is a mental unit. However, in concept theory, the definition of concept is a knowledge unit. Dahlberg in her theory disagrees that the concept can be a mental unit, because such understanding can vary from person to person; according to the mind of each person, it is necessary to have an agreement between people to allow communication. Hence, it defines concept as a unit of knowledge, where an ontological commitment is established between the members of a given community.

Through the analysis of the faceted classification theory and the concept theory, it is clear that these theories support one of the first methodological efforts, in information science, to highlight the representation unit issue. Ranganathan, in his Prolegomena, introduces the concept of “isolated” as the classification system’s minimum and manipulative unit. Dahlberg (2014) presents the “concept” as a minimum unit and defines it as a triad, composed of the elements referent, characteristics, and name. In this triad, the object is the referent, classified as an individual or general
object that, when circumscribed to a context, requires the appropriation of characteristics, being designated a linguistic sign—a name.

In the context of formal ontology, in computer science, Guarino (1998a, 1998b) proposes that objects or “particulars” be classified as concrete and abstract. Comparatively, in concept theory, these are considered general objects, as they represent a class of objects, not a particular entity in the world. Concrete objects, within the scope of formal ontology, are classified as continuous and occurring. These discussions aim to highlight the multiplicity of interpretations and end up allowing flexibility for observation and work in different domains. In addition, they make the modeler feel the need to build the assumptions where to start from or even try to understand where he starts from in order to develop a representation model.

### 2.1.4 Relations between objects

The fourth principle has the purpose of verifying the possibilities of semantic connection/separation between the concepts of a domain. The relationships between objects within a context form the conceptual structure of that context and are of different natures. The analysis is presented in the movements of the act of modeling. These movements reflect groups of relationships between concepts, as described, in an exemplary manner, below.

The first movement is the verification, in the presented theories and methods, of the existence of categorial relations. This type of relationship brings together, in a first major grouping, objects by their nature, that is, entities, processes, among others. This relationship often makes it possible to reduce logical errors in establishing links between concepts, as it determines the nature of the object.

Having verified the reunion or not of objects by nature, the second movement, not in order of precedence, but of necessity, is to verify how objects of the same nature are related. In this form of relationship, there is a determined hierarchical relation. Consequently, theories and methods present relations that determine “what is the object,” the understanding of the concept. Accordingly, the theory of the concept and the theory of classification call this relationship hierarchical; the general theory of terminology calls it a logical relation, and formal ontology defines this relation as “one (ISA),” considered a property in a minimal ontology of universals for the structuring of a domain.

Another movement that implies the relationships between the concepts is the analysis of how “the object is constituted,” that is, its parts and elements. In this form of relationship, partitive relations are determined. Like the other relationships, this one is named differently in the theories and methods analyzed: 1) the concept theory is called partitive relationship; 2) the classification theory places hierarchical and partitive relations in the same group, calling them hierarchical relationships; 3) in terminology theory, it is called the ontological relationship of partitive coordination; and, 4) in formal ontology, partitive relations are named all-part relation, also called “mereotopology.” Formal ontology presents more explicit criteria for determining the type of partition. These criteria help the modeler to distinguish the nature of what will be considered as part and as a whole and function as “informational meta-categories” to reflect on the relations of a domain.

The next movement is to see how objects of different natures relate to each other and to represent that relationship more consistently. That is, in this process, some prescriptive criteria are determined that allow for more judicious connections. The relationship between concepts of different natures can be observed in the concept theory as a syntagmatic functional relationship. These relationships, unlike paradigmatic ones (logical and partitive), can be recognized as relationships that make evident a specific demand or function between objects in the phenomenal world and do not aim to explain the object and its properties.

Finally, another movement to address no longer has to do with comparing or the relation between concepts but between how to express these concepts. In other words, this movement happens in the language sphere, the so-called equivalence relation. Once again, a space for analyzing these relationships builds on the different theories and methods that encompass the thematic scope of domain modeling.

### 2.1.5 Forms of representation

This principle contemplates the forms of graphic representation that a model can adopt. A conceptual model must also be a communicational space where the phenomenal world transposes into a space of representation. Therefore, it is important to investigate graphic forms of representation, because sometimes these forms impair representational possibilities. In the theories and methodologies analyzed, it appears that the area of information science, despite having well-grounded theories about the concept and conceptual relationships, is scarce in terms of models that assist in the development of graphic representations. In the terminology context construction practices, where there is a configuration of general theory of terminology and ontology, graphic representations are consistently present, although there are some fundamental differences.

Even if graphic representations are not addressed, it is often, in some areas of knowledge or professional fields, important to mention that they lack, much more than any other point analyzed, a space for investigation. It associates the relevance of studying them with their contribution by making the twinned work processes in the representation of knowledge more transparent. In corporate education,
example, these representations may contribute to improvements in the management of learning paths. Next, we will discuss the representation of learning paths in a corporate education context, where we consider the theories discussed in this section. In this sense, as a conclusion of the principles exposed here, we present the summary table of the principles for domain modeling (Table 1). In the summary table, we will highlight the theories that we discussed for the reasoning method, for the identification of the unit of representation, and for the relationships between objects.

### 3.0 Representation of learning paths in the context of corporate education

According to Lopes (2018), learning paths can be understood as a systematic and multimodal set of learning units, containing different navigation schemes. These schemes can range from linear and prescriptive models to more hierarchical models and to network models, where navigation is freer and with the purpose of developing skills. The navigation schemes in learning paths can be customized, based on variables such as objectives, student profile, and learning characteristics.

In the education field, a learning path is fundamental to the teaching-learning process, since it integrates a set of activities in an appropriate sequence, allowing the student to learn the contents more effectively. A learning path, according to Yang (2012), comprises different activities aimed at learning, and these activities can relate to different teaching approaches, seeking whenever is possible a learning path that is closer to the performance and preferences of the student.

From the librarianship and information science’s view, we can analyze the learning paths and hypertext systems similarly, as highlighted in the study by Lopes (2018). To a certain extent, in Campos’ (2001) view, hypertexts can be compared to a system of concepts, as they are conceptual nodes linked to other nodes through a level of relationship. Campos (2001) points out that, coupled with “being hypertextual,” there is a classificatory action since the connections between the nodes implement from a network of associations.

The existence of a route in a learning path, whether previously shown or created by the learner himself, is close to the hypertextual relationships that can manifest within the

<table>
<thead>
<tr>
<th>Foundational theories of Information Science</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Principles</strong></td>
</tr>
<tr>
<td>1. Reasoning Method</td>
</tr>
<tr>
<td>2. Object/Unit of Representation</td>
</tr>
<tr>
<td>3. Relations between objects</td>
</tr>
<tr>
<td>3.1 Categorical Relationship</td>
</tr>
<tr>
<td>3.2 Hierarchical/Logical Relationship</td>
</tr>
<tr>
<td>3.3 Partitive Relationship</td>
</tr>
<tr>
<td>3.4 Relationship between objects of different natures</td>
</tr>
<tr>
<td>3.5 Equivalence Relationship</td>
</tr>
<tr>
<td>4. Forms of Graphic Representation</td>
</tr>
</tbody>
</table>

Table 1. Summary Table of principles for domain modeling.
In corporate education, the educational solutions organization learning paths is a reality in many institutions since this type of strategy can contribute to the processes of planning, execution, and evaluation of training actions as explored in the researches advocated by Brandão (2012), Carbó (2013), Freitas and Brandão (2006), and Murashima (2011). The results revealed, in general, that learning paths can be organized by areas, positions, competences, and work processes, among other categories.

In this section, based on the five fundamental principles for the act of modeling knowledge domains, discussed in Section 2, it is sought, based on studies from the fields of information science, computer science, and terminology, to verify in literature the extent to which these principles are applied. Furthermore, as a concept proof, a conceptual model of learning paths is elaborated, applying the determined principles.

3.1 Application of the principles of domain modeling in the analysis of learning paths construction methodologies

In this section, we seek to present the methodological procedures adopted in the comparative analysis carried out between the domain modeling principles mentioned in Section 2 and the strategies for building learning paths presented in the scientific literature in the area of corporate education. The research strategy adopted was exploratory and qualitative. To achieve the proposed objective, three stages of data collection and analysis were established, considering the content analysis method proposed by Bardin (2011), namely: 1) pre-analysis; 2) exploration of the material; and, 3) treatment, inference, and interpretation of results.

First, in the pre-analysis stage, we carried out an exploratory reading of the literature review publications (Lopes and Lima 2019). Subsequently, we selected forty-three publications as a corpus of analysis considering as a criterion the studies whose authors explained the stages of the conceptual modeling process. We assume that in the publications on organization and representation of learning paths, the authors possibly did not maintain adherence to the principles of the conceptual models of hyperdocuments proposed in Section 2 of this article. In the stage referring to the exploration of the material, the KOS knowledge domain modeling principles addressed in this study are categories of analysis.

Although Bardin (2011) shows an inductive logic for establishing the categories of analysis, starting from the coding from the exploratory reading, the deductive logic was chosen, since the principles presented in Section 2 were representative and convergent to the objective established for this stage of the research. Thus, for the organization and coding of the data, we structured a matrix, contemplating the referred principles, and the publications selected analysis corpus. For each publication, there was a new reading, and each principle’s presence or absence inserted in the matrix. For each principle, there were subcategories of analysis related to each of the principles associated with the modeling of knowledge being raised.

Finally, in the stage of treatment, there was an inference and interpretation of results, a comparative analysis of publications and each principle was identified. It should be noted that, although they have different characteristics, each step was performed in an integrated manner in view of enabling the complementation of the results from the literature, which are presented below.

Through the analysis of the “domain of knowledge to be modeled” principle, we consider three subcategories of analysis, namely: 1) if there was a paradigmatic cut; 2) if there was a syntagmatic cut; or, 3) if it was not possible to identify the type of thematic cut adopted. In 39.5% of the publications, the authors explain a paradigmatic outline, focusing mainly on disciplines of academic curricula. To a lesser extent, in only 13.95% of publications analyzed, the authors predicted syntagmatic cuts. As explained by Campos (2001), when a paradigmatic cut is identified, the knowledge units are rigorously hierarchized. On the other hand, from the syntagmatic cut, the knowledge units form different hierarchical trees, brought together globally. In this sense, we observed that the syntagmatic cut meets the interdisciplinary and polychotomous nature of the themes identified in an organizational context, pointing to the formation of complex networks of knowledge units. Finally, it was evident that most publications in the analysis corpus (46.5%) did not foresee or did not explain how the thematic cut of the content was done. There is an assumption that this result may have been influenced by the different focuses of publications, which do not necessarily explain the adopted methodology.

Through the analysis of the “reasoning method” principle, we envisaged the following subcategories of analysis: 1) application of the deductive method; 2) use of the inductive method; 3) use of the analytical-synthetic method; and, 4) no possibility of identifying the reasoning method. In 41.86% of the publications in the analysis corpus, the authors explained the application of only the deductive reasoning method. Publications that address the inductive method start from lists of disciplines, themes, learning objectives, courses, and skills to be developed. Specialists determine these lists previously, and generally, they are linear or hierarchical. There is also a need to highlight the prevalence of prior validation of conceptual schemes by specialists in the thematic domain, characteristic of deductive approaches. Regarding publications that explained only the inductive method, in 37.2% of the publications in the analysis corpus, the authors are directed to this...
focus. Regarding the analytical-synthetic reasoning method, only 13.95% of publications can be classified as such. Finally, in only 6.97% of publications, it was not possible to identify the reasoning method adopted.

In the analysis of the “definition of the representation object” principle’s view, publications that mentioned the requirements of: 1) granularity; 2) descriptive representation.; 3) unit of knowledge be self-contained; and, 4) publications that did not mention this principle. In the scope of this study, the concept of granularity is the smallest unit of information. The subcategory “descriptive representation” refers to so-called linguistic labels or linguistic signs used to represent units of knowledge. Regarding the quality of a self-contained knowledge unit, it relates to the importance of defining the scope of content treated in the so-called conceptual node, seeking to contemplate what the concept is and its constituent elements. In the English language, there is a specific term to refer to this quality, which would be the term self-contained. The data revealed that 27.90% of publications mention granularity, while 39.53% address descriptive representation. While in only 13.95% of publications, the authors refer to questions of granularity and descriptive representation together. Only one identified publication referred to the importance of the knowledge unit being self-contained. Finally, 39.53% did not explicitly indicate this principle. Based on the analysis of publications that did not mention the nature of the knowledge units’ content, it was possible to verify that, generally, the proposals for modeling learning paths are restricted to the representation of categories and subcategories with a more global and not so specific focus that could contemplate the nature of the content. Regarding the granularity item, we observed that there is no standard as to the determination of which is the smallest node on the learning paths. Regarding the descriptive representation, there was also a lack of clarity regarding the standards adoptable for the different documentary typologies. Finally, in relation to the item “to be self-contained,” in only one publication was explicit reference made that the knowledge unit is self-contained, which could favor reuse and interoperability.

In view of the analysis of the “relationships between objects” principle, the types of relationships provided as subcategories of analysis, namely: categorical, hierarchical, partitive, between categories, and equivalence. It was also considered as a subcategory of “publications whose principle was not addressed by the authors.” A hierarchical relationship was identified in 51.16% of the publications, while in 83.72% there is a relationship between different categories, considering that in 46.51% of the publications the authors mention two or more types of relationships. The categorical and the equivalence relation had only one mention each. The partitive relationship was never mentioned. In only three of the publications, the authors did not explicitly describe the establishment of relationships between knowledge units. Among the publications whose authors explain relationships, these can be established a priori, by a specialist in the field, and by the student himself, according to learning objectives and preferences. Relationships can also establish automatically, considering several variables such as prior knowledge, learning style, performance, student profile, and learning objectives, among others. We emphasize the identification of “ontological chaining” relationship in which the learning path’s construction considers the predecessor and successor units of knowledge. There is also an emphasis on the application of data mining techniques for the automatic generation of learning paths.

From the analysis of the principle “forms of representation,” it was possible to identify the following analysis subcategories: concept map, flowcharts, business process management notation (BPMN), mind map, website or system, and publications that did not address this principle. The representations in the flowchart formats were identified in two publications; there was only one mention of the BPMN notation and mind map, each. In 11.62% of the publications, the authors covered both websites and systems and concept maps, each. We also identified that in 69.76% of the publications, the authors do not clearly mention the graphical representation of the nodes and relationships from the point of view of the information visualization by the user. However, in 46.51% of the publications, the authors explicitly address some form of representation of the conceptual model adopted for the learning paths. Of these representations, the ontologies, taxonomies, and architectures of the systems were used to implement the learning paths.

Through comparative analysis, the results revealed, in relation to the “domain of knowledge to be modeled” principle, that there is adherence to this principle in the methodologies for building the learning paths identified in the consulted literature. However, there was partial adherence, since almost half of the publications (46.5%) do not mention this principle. It is estimated that the focus on the paradigmatic outline does not portray the polytomous nature of the learning paths, which can be composed of knowledge units from different domains, not being restricted to a canonical area.

Regarding the “reasoning method” principle, considering the recommendation to adopt the synthetic-analytical method in order to privilege a systemic perspective for the organization of knowledge units, it was observed that there is low adherence to this principle in the identified methodologies. This finding is justified, since in only 13.95% of the publications the authors made explicit the adoption of an analytical-synthetic approach.

Regarding the “definition of the object of representation” principle, 60.47% of publications were identified. However, the data revealed that the lack of a standard for
determining granularity and documentary typology may compromise the possibility of reusing the units of knowledge provided for in a learning path.

In relation to the principle “relations between objects,” adherence was identified, considering that 93.03% of publications mention some sort of relationship between knowledge units. This result was verified, even considering that it was not possible to identify among the publications any whose authors addressed a clear orientation regarding the description and possibility of applying each type of relationship. Consequently, it is a research opportunity to provide guidance on these relationships’ types and to apply them in a learning paths context. Regarding the principle “forms of representation,” there was adherence, since 79.05% of publications explicitly address some graphic representation of the conceptual model or some form of learning paths visualization, however, only from the users’ point of view.

In general, the principles for modeling knowledge domains presented in Section 2 are applicable to the context of creating methodologies for the organization and representation of learning paths. In this specific context, the inclusion of a principle focused on data storage and management systems is an expandable opportunity, considering the current society’s scenario, sustained by a large volume and diversity of information. The application of these principles, as a concept proof, in the formulation of a conceptual model is presented below. Such a model can be understood as the manifestation of a graphic representation for a better visualization of the knowledge representation processes.

3.2 Proposal of a conceptual model for organization and representation of learning paths

The conceptual modeling stage is essential in the knowledge representation process. However, in the context of creating hyperdocuments, it is not a stage always contemplated. The proposal for a conceptual model aims to assist the authorship of hyperdocuments in order to establish a communication space whose content author and other professionals involved in the process of creating the hyperdocument can interact and jointly represent the conceptual model (Campos 2001). In this sense, a model is proposed for the organization and representation of the learning paths, especially for the area of corporate education. This model allows exemplification of the modeling principles presented in this study, and that is applicable in other contexts.

Regarding the “domain of knowledge to be modeled” principle, there are two structure possibilities, both under a paradigmatic cut and under a syntagmatic cut. In the first possibility, the knowledge units are part of the same canonical area. For example, in learning paths on librarianship and information science, the tendency is for knowledge units to group in this knowledge field or domain. In the second possibility, knowledge units have more than one subject area. For example, a learning path composed of knowledge units from a business process and units of knowledge from a canonical area.

About the “reasoning method,” in the organization of knowledge units that will compose learning paths, it is possible to adopt deductive, inductive reasoning methods, or even the analytical-synthetic method. When using the deductive method, the general categories that will comprise the units of knowledge must be pre-defined. These general categories should be exhaustive, with the capacity to accommodate all units of knowledge that intend to be covered within the learning paths coverage. In the inductive reasoning method, it starts from the analysis of knowledge units and their respective attributes, looking to identify similarities that may indicate categories of subjects that allow the grouping of these knowledge units. Finally, but not least relevant, in the analytical-synthetic method, constant movements of analysis and synthesis are carried out, using the deductive and inductive methods in an integrated manner. The analytical-synthetic approach is the most appropriate for the context of the learning paths organization.

Concerning the “definition of the object of representation,” in theory, this object in the context of learning paths is the competence intended to develop. This competence represents a concept. Therefore, the smallest unit of manipulation/representation in the context of learning paths is the concept. Based on the premise that learning is a process, the representation object can be classified as an abstract object, occurring, or a process within the formal ontology’s scope or else an action in the concept and terminology theories contexts.

Regarding “relationships between objects,” starting from the movements proposed in Section 2, it was possible, through the proposition of a conceptual model of organization and representation of learning paths, to identify categorical relationships. These categories are like those proposed by Ranganathan in the theory of faceted classification or even those suggested by Dahlberg in the theory of concept.

In order to identify the way in which knowledge units relate to each other on learning paths, we found that the hierarchical relationship is essential, and it is present at different times in the organizing process of learning paths. In this context, we seek to identify the nature of the knowledge unit, providing for analyzes on how these educational solutions are hierarchized, in which types of categories these solutions are classified, among other possible analyzes. In a subsequent movement, the aim is to identify which parts may compose the same educational solution. A learning path is formed by different educational solutions, composed of learning objects, learning activities, and assessment activ-
A learning object can form from different units of knowledge, represented by videos, texts, and podcasts, among other forms of representation. In a complementary way, in the fourth movement, we seek to verify how objects of different natures relate and to represent this relationship in a more consistent way. In this sense, we identify the possibility of relationships between these objects, for example, in the relationship between an educational agent (tutor, teacher, mentor, etc.) and a given educational solution (course, video, seminar, tutorial, etc.). In the fifth and last movement, to express the different units of knowledge that may compose a learning path, there is a use of linguistic signs that enable equivalence relations between different units of knowledge.

Finally, the “forms of representation,” identifies that the learning paths may be represented, graphically, by a simple or hierarchical list of knowledge units, mind maps, concept maps, specific notations, or graphs, among others. As in information science, graphic representations of learning paths also demand greater space for investigation. With this, it becomes possible to identify forms of representation of the knowledge units that mainly contain the diversity and the volume of possible units of knowledge to compose learning paths.

In Figure 1, the conceptual model is presented, as concept proof, for organization and representation of learning paths, structured based on the modeling principles of knowledge domains described in Section 2.

The model presented in Figure 1 is organized into four major categories: actors, components, properties, and processes, showing, in the first level, the relationship of the categorial type. In each of these major categories, same nature concepts are grouped. In the first category (actors), the grouped concepts have a hierarchical coordination relationship. In the second and third categories (components and properties, respectively), the partitive relationship is identified, considering that the components and properties are part of learning paths. In the fourth category (processes), there is a relationship of the hierarchical type of coordination. In the conceptual model, we opted to hide part of the predicted subcategories, considering the focus of this work, which is on the knowledge domains’ modeling and also due to the space limitation.
4.0 Final considerations

This study addressed a set of principles for the development of domains’ conceptual models and the verification of the application of these principles in a set of methodologies for building learning paths. We identified the methodologies in a literature review process, and, as a proof of concept, we proposed another conceptual model of learning paths based on the identified principles. It was observed that the principles for modeling knowledge domains addressed are partially applied in the methodologies for building learning paths identified in the consulted literature. Starting from the assumption that each proposed principle bases itself on fundamental theories in the areas of information science, computer science, and terminology and also considering that such principles have yet to be fully applied in the modeling processes of knowledge domains, as we presented in this study, one of our findings is that there is an opportunity for greater dissemination of these principles among professionals working in the modeling of knowledge domains.

Based on guiding principles, the professional who works in the modeling of knowledge domains has the possibility of overcoming pre-established models, being able to understand the metaconceptions involved in the modeling process, used as abstraction mechanisms. This observation is aligned with the reflection proposed by Ranganathan, still in the twentieth century, on the importance of having more classificationists, those who elaborate classifications, and not just classifiers, those who use classifications.

Considering that the application of the principles of knowledge domains modeling addressed enabled the systematization of a conceptual model, the application of these principles in the field of corporate education, especially in the construction of learning paths, was of fundamental importance. Certainly, these principles contribute to the quality of the results achieved in addition to the possibility of proposing a new model for building learning paths. Therefore, it is possible to assume that the principles discussed in this study can be considered relevant as they enable the modeler’s freedom, not leaving him hostage to a specific model. Thus, the application of the proposed principles in other fields of knowledge is expected, aiming the creation of new proposals for meta-representations that can contribute to the processes of knowledge organization and representation.

References


Representing and Linking Dunhuang Cultural Heritage Information Resources Using Knowledge Graph

Xiaoguang Wang*, Wanli Chang** and Xu Tan***

Wuhan University, School of Information Management, Wuhan, Hubei, 430072, CN,
*<wxguang@whu.edu.cn>, **<wanlichang@whu.edu.cn>, ***<txishuo@whu.edu.cn>

Xiaoguang Wang is the vice dean and a professor in the School of Information Management at Wuhan University, China. His research interests include digital asset management, knowledge organization, semantic publishing, and digital humanities.

Wanli Chang is a postgraduate student in the School of Information Management at Wuhan University, China. His research interests focus on data management of cultural heritage and knowledge discovery.

Xu Tan is a doctoral student in the School of Information Management at Wuhan University, China. Her research interests include digital humanities, semantic web technology, digital libraries, and cultural informatics.


Abstract: This study employs a knowledge graph approach to realize the representation and association of information resources, promote the research, teaching, and dissemination of Dunhuang cultural heritage (CH). The Dunhuang Mogao Grottoes is a UNESCO world CH site, and digitization of Dunhuang CH has produced a large amount of information resources. However, these digitized resources continue to lack the systematic granular semantic representation required to correlate Dunhuang cultural heritage information (CHI) in order to facilitate efficient research and appreciation. To respond to this need, new approaches for representing CHI are being developed. This study identifies five facets and their semantic relationship to Dunhuang CH, constructs an ontology model to regulate the entities, attributes, and relationships of Dunhuang CH knowledge, and subsequently processes the resulting data using various techniques (such as semantic annotation and entity association) to facilitate rendering the data in a knowledge graph construction. Finally, we constructed a DH-oriented knowledge graph service platform in order to provide a user-friendly visual display and semantic retrieval service.

Received: 1 June 2020; Revised: 30 September 2020, 16 October 2020, 23 October 2020; Accepted: 23 October 2020

Keywords: knowledge, Dunhuang, data grottoes semantic

1.0 Introduction

Cultural heritage (CH) resources are numerous and widely distributed, containing rich and diverse cultural knowledge. Effective protection and utilization of CH resources is of great significance to the sustainable development of human civilization. Due to the growth of digital humanities (DH), the acquisition, organization, management, and dissemination of CH resources using digital technology has aroused great interest from libraries, archives, and museums (LAMs) in recent years. In the field of Dunhuang CH, the development of digital technology has produced fruitful results over the last forty years as evidenced by the proliferation of digitized Dunhuang cultural heritage information (CHI) resources.

Structuring and storing information resources in databases is the basic method currently used to facilitate the management and organization of CHI. Through the implementation of knowledge organization system (KOS) principles and best practices, fine-grained and semantic description of CHI can be used to facilitate knowledge management and mining. KOS is a universal term for the various
norms and methods used to organize knowledge. It is an important means of acquiring and utilizing knowledge and is a generic term for semantic tools used to express and effectively organize the elaboration of various knowledge structures. KOS is divided into four sub-levels based on the complexity of the structure and main functions, namely: term lists, metadata-like models, classifications and categorizations, and relationship models (Zeng 2010). In the field of CH, the semantic representation and linking of CHI can be achieved by using different levels of KOS.

A series of thematic databases have already been established in the field of Dunhuang CH. Such as, the International Dunhuang Project (IDP), Mellon International Dunhuang Archive (MIDA), E-Dunhuang, Dunhuang Academic Resources Database, and so on. These projects have organized a plethora of DH resources and improved public access to them. However, many impediments to efficient scholarship and appreciation of Dunhuang CH still persist today. From a KOS perspective, these shortcomings include:

1. Metadata and taxonomy are used by these discrete databases to provide a basic description of their information resources. However, this basic flat description is no longer adequate in today’s multi-source, heterogeneous, multimodal, big data environment, where data can also be semantically described and linked through more advanced knowledge representation and organization models (such as ontologies and knowledge graphs) for the purpose of deep revelation of resource content. In addition, problems also persist in databases management and interoperability posed by redundancy in the construction of resource repositories employing different elements and properties in these flat resource descriptions.

2. Although there are a lot of images and 3D resources residing within these databases, the cave structure of Dunhuang CH might not be accurately described in the absence of related data models.

3. In the context of DH, data service providers need not only provide digital information resources but must also integrate the information needs of users to provide associated resources and context. However, most existing Dunhuang related databases lack this necessary connection to external resources and other databases are document-centric rather than entity-centric. This approach has limited the value and usability of these discrete information silos.

Therefore, this study establishes the Dunhuang Grottoes Knowledge Graph and user-oriented application service platform covering multimodal resources with functions such as semantic retrieval and knowledge navigation to provide better support services for the protection, research, teaching, and dissemination of Dunhuang CH. Such a model would be capable of accurately and comprehensively revealing intrinsic features and external information regarding Dunhuang CHI and finally meet the needs of knowledge mining and analysis in order to revolutionize the efficiency of knowledge utilization.

2.0 Related work

Semantic representation and organization of CHI is the basis for open sharing and knowledge mining. At present, communities of LAMs have developed many general and special KOSs (knowledge organization system). For example, the metadata models include DCMI (Dublin Core Metadata Initiative), VRA Core (the core categories for visual resources), CDWA (Categories for the Description of Works of Art), RDA (Resource Description and Access). The thesaurus includes AAT (Art and Architecture Thesaurus), ULAN (Union List of Artist Names). The ontologies include CIDOC CRM (CIDOC Conceptual Reference Model), ABC, FOAF (Friend-of-a-Friend), and so on.

In accordance with the growing trend towards big data in CH, massive multi-source heterogeneous data must be organized and linked, and a knowledge graph approach provides the possibility for this. A new knowledge representation technology of KOS—knowledge graph has ontology technology at its core. As a result, knowledge graph provides a path and method for extracting structured knowledge from multimodal digital resources and representing and storing this structured knowledge in a graphical data model (Singhal 2012). With the help of knowledge graph, advanced forms of knowledge services such as knowledge visualization, semantic retrieval, and intelligent question answering can be better realized. The core technologies of knowledge graph mainly include entity relationship identification and extraction, knowledge fusion, entity linking, and knowledge reasoning (Moura et al. 2014). Moreover, knowledge graphs are widely used in intelligent search at well-known internet companies, such as Google (based on Wikidata and freebase database), Microsoft (Satori Knowledge Base), and Baidu. With the development of knowledge graph technology, knowledge graphs have also been successfully applied in CH projects, such as Europeana (Purday 2009), Getty Digital Museum (Alexiev et al. 2015), Venice Time Machine Project (Abbott 2017), and Finland’s CultureSampo project (Mäkelä et al. 2012).

This study contends that knowledge graphs are also particularly useful in linking and organizing the rich and inherently complex CHI of three-dimensional structures. Historical buildings are valuable and a non-renewable CH resource. In order to preserve, restore, and recreate historic buildings, an increasing number of research projects are employing digital technologies to describe, manage, and ana-
lyze architectural heritage information. 3D modeling is an important technology in CH conservation, which primarily uses computers to describe spatial relationships between objects in a mathematical way. Allen et al. (2003) performed a comprehensive 3D modeling of St. Pierre’s Cathedral using a 3D laser scanner, and Foni et al. (2007) of the University of Geneva use the Cathedral of Saint Sophie as an example in order to perform a virtual reconstruction and simulation of a complex and endangered historical building.

In the field of architectural engineering and construction (AEC), building information modelling (BIM) is widely studied and applied with regard to the development of digital cities and smart cities. BIM is the integration of the physical and functional information of the whole life cycle of the building, which contains rich and detailed structural information. Due to developments in interdisciplinary research, some DH scholars have begun to use architectural information models to represent and organize the structural information of historical buildings in the field of CH. For example, Simeone integrates building information systems and ontology-based systems to enhance knowledge representation and management in the process of building heritage (Simeone et al. 2019). Similarly, Noor et al. (2018) believes that the traditional architectural heritage metadata model standard lacks a description of the key aspects of 3D modeling and proposes a model that uses semantic web technology and building information models to represent building data and historical data. Pauwels et al. (2013) also achieves the integration of architectural elements and building-related CHI, including historical documents, design drawings, etc., by combining building information models and semantic web technologies to demonstrate the feasibility of this approach using the Ghent Book Tower in Belgium as an example.

3.0 The semantic contents of Dunhuang CH

The triadic world refers to the spatial form made up of the physical world, human society, and information space, the connectivity and integration of which is an important feature and trend of the new generation of computing technologies (Ma et al. 2019). In the context of DH, with the continuous development of digital technology, a large number of digital resources have been generated on the basis of CHI. Traditional humanities research is gradually shifting to DH research, and this evolution in research paradigm has gradually transformed the structure of knowledge space in the field of CH from the original binary space (i.e., physical space and cultural space) to a ternary space. According to the triadic world theory, Dunhuang CH can be divided into physical space, cultural space, and information space: 1) physical space refers to the physical space constituted by the internal architecture of the Dunhuang grottoes and their external rock walls, which is an external manifestation of Dunhuang CH; 2) cultural space refers to the mythological world and human social world depicted by the Dunhuang grottoes. (the mythological world generally includes Buddhist figures narratives and time, as well as other religious figures, non-religious mythological figures, and so on. The human social world covers elements, such as administrative institutions, religious organizations, general social groups, ancient ancestors, and folklore); and, 3) the information space is a virtual space that realizes the digital twin of Dunhuang CH physical space and cultural space through the flow of information using the Dunhuang CHI element as its carrier.

In consultation with Dunhuang Academy research scholars, this paper divides the physical space, cultural space, and information space of the Dunhuang grottoes into five facets. These facets include: cave structure knowledge, academic works knowledge, digital information resource knowledge, mural content knowledge, and, finally, the facet of historical and cultural knowledge. Each of these facets is discussed in further detail in the following section.

3.1 The facet of cave structure knowledge

A religious product, grottoes provided a mechanism for transmitting Buddhist ideas to subsequent generations. Therefore, they contain rich historical and cultural information. The different forms of grotto architecture express not only the development of Buddhism but also the shift from one cultural form to another. Therefore, they are an integral part of the process of organizing CHI.

The Dunhuang grottoes are a type of grotto temple building constructed on the cliffs of the Gobi Desert in northwest China, where the design of the interior space is more important due to the influence of the external environment. The interior of Dunhuang grottoes is a closed structural space. There is a commonality between different grottoes. No matter how the internal structure of the main room changes, the grotto always maintains the same basic spatial pattern composed of three parts: the front room, the main room, and the hallway. Moreover, in addition to the east wall with doors or windows, there are many beautiful murals painted on the north, south, and west walls as well as on top of the caves. For example, Cave No. 285 displays a classic grotto temple structure (Figure 1). The architectural form of the main room of the grotto may be subdivided into four categories. These categories include: the zen grotto, the central pillar grotto, the overlying grotto, the back-screen grotto, as well as a small number of big Buddha grottoes, nirvana grotto, and altar grotto (Ning 1986). The variations in the forms of each of the Dunhuang grottoes
indicates the architectural structures of different grotto forms were determined by their historical and cultural functions.

3.2 The facet of academic works knowledge

In the process of studying the history and culture of Dunhuang, Dunhuangology was formed. Dunhuangology is practiced by many academic fields, including: religion, art, history, archaeology, language, literature, anthropology, geography, philosophy, science, technology, architecture, ancient book collation, cultural relics protection, sports, folk custom, and so on. Over the past century, Dunhuangology has been continuously developed and deepened. The study of Dunhuang grottoes has been very fruitful, as a result, a large amount of academic research resources, including research papers, books, archaeological reports, and picture albums have been created. These academic works and documents have provided important value for the excavation of CHI in the Dunhuang grottoes. Fan (2000) divides the stages of research into the Dunhuang grottoes into three periods: beginning, development, and full development. From the beginning of data collection, picture interpretation, content verification, to topic discussion, comprehensive research, etc., a great deal of research and analysis has been published. These studies have laid a strong foundation for further in-depth research in the future.

3.3 The facet of digital information resource knowledge

Digital information resources are an important manifestation of CHI. They can reshape the physical space of the grotto through digital photography, three-dimensional modeling and other digital technologies, allowing the grotto to break through its physical and geographical limitations. The characteristics of digital information resources can be divided into data characteristics, structural characteristics, and content characteristics. Data characteristics that describe that CH digital resources are massive, multi-source, heterogeneous, multi-modal, and cross-cultural. Structural characteristics describe the medium or format of CH digital resources, including audio, video, photographic, and so forth. Finally, content characteristics include digital murals, digital archives, 3D models, laser point clouds, multi-spectral images, etc. At present, the Dunhuang Academy has cooperated extensively with renowned scientific research institutions and universities to form a set of technical specifications for the digitization of cultural relics suitable for the characteristics of immovable cultural relics such as grottoes. It has accumulated several digital resources of different data types, including multimedia resources such as pictures, audio, video, and 3D models.

3.4 The facet of mural content knowledge

The artistic composition of the Dunhuang grottoes has four main elements: grotto temple architecture, statues, frescoes, and Dunhuang MS. The most important part of Dunhuang CH, Dunhuang murals are cultural treasures in the history of mankind. Considered to be magnificent “libraries on the walls,” they have extremely high artistic and appreciation value. Because the murals display the creativity and integration of Dunhuang artists of all dynasties, they formed their own unique artistic characteristics and techniques for expression. Compared with other elements, the
muralsthe Dunhuang Grotto Ontology Model was extended and customized to the characteristics of Dunhuang grottoes CH by referring to the widely used and recognized ontology models in the world.

The Dunhuang Grotto Ontology Model displays a high level of abstraction and generalization of Dunhuang grottoes domain knowledge. It expresses informal knowledge clearly as various concepts and their relationships. It deeply describes and reveals the semantic characteristics of Dunhuang grottoes’ research resources, and the semantic connections between them. According to the division of Dunhuang CH by the ternary world theory, the knowledge of Dunhuang CH may be divided into five facets. Moreover, knowledge of Dunhuang’s CH is organized around the Dunhuang grottoes (the structure of grottoes and the contents of its murals) on the basis of “people-things-time-land-things.”

The Dunhuang Grotto Ontology Model references and reuses CIDOC CRM, FOAF, FRBR and other internationally widely used ontology models and data standards. In addition to this achievement by the Dunhuang Grotto Ontology, this paper formulates corresponding data standards and data mapping specifications for the literature, multimedia resources, databases, metadata information, and other data in the field of Dunhuang grottoes in order to finally provide a unified and standardized data model for the construction of a knowledge base for the Dunhuang grottoes. In accordance with the core constituent elements of Dunhuang grottoes CH, the data model is divided into the following classes: grotto, indoor space, component, clay sculpture, mural, subject, academic work, digital resource, event, location, time, and agent. Each class is associated and described by object attributes, and each class has its own defined data attributes.

4.1 Modeling of the spatial structure

When expressing the spatial structure in the Dunhuang grottoes, the existing international standards such as CityGML and IFC of BuildingSmart provide the geometric expression, attributes, and visual data model of the building composition. However, these standards do not provide for detailed description of the orientation relationship inside the indoor space structure. This inability results in the loss of some information in the process of knowledge organization (Kim et al. 2014). Moreover, although these model standards can describe modern buildings well, they fail to meet the need for fine-grained semantic representation in the field of CH architecture. In particular, different CH buildings have very different architectural characteristics, which needs to be taken into consideration when organizing structural knowledge of CH buildings.

The spatial structure of the Dunhuang grottoes can be semantically divided into indoor space, components, and indoor objects according to the characteristics of the spatial structure of the grottoes (Figure 2).
Semantic location is an important means of linking semantic objects in space and can be direct or indirect. Direct semantic positional relationships include: coordinates, place names, addresses, etc., while indirect semantic positional relationship may include a combination of distance, orientation, and relative topological relationships. In this paper, the semantic objects are mainly related through indirect positional relationships. Reference is made to the division of spatial positional relationships in MPEG-7, which are spatial orientation relationships, topological relationships and spatial semantic relationships. These spatial position relationships are discussed further in Table 1 below.

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial direction relationship</td>
<td>Represents regional location information of semantic objects in space</td>
<td>top of, bottom of, middle of</td>
</tr>
<tr>
<td>Topological relationship</td>
<td>Spatial orientation location relative to other components of or in the space.</td>
<td>nearby, within, contain, adjacent to</td>
</tr>
<tr>
<td>Spatial semantic relationship</td>
<td>Represents the logical relationship between semantic objects in space, such as subordination, dependency, etc.</td>
<td>belongs to, part of, consist of</td>
</tr>
</tbody>
</table>

Table 1. Spatial position relationship.

According to the semantic division of the spatial structure and the classification of spatial orientation relationships, this paper designs a spatial structure ontology representation model suitable for accurately representing the Dunhuang grottoes. This new spatial structure ontology representation model includes five categories: grottoes, space, components, statues, and murals. The data attributes of spatial orientation are defined for each class, signifying the spatial orientation relationship, topological relationship, and spatial semantics between the cave structures.

4.2 Modeling of the academic work

Academic research in the area of Dunhuang grottoes has accumulated a large amount of academic research resources, including research papers, books, and paintings. These academic works and documents provide important documentation and analysis of the historical and cultural knowledge contained in the Dunhuang grottoes. The facet of academic work knowledge provides a correlation between research literature and the content of Dunhuang grottoes in order to reveal and link the academic knowledge contained in the field of Dunhuang grottoes. The attributes of the facet academic work knowledge mainly include the titles, authors, introductions, editions, publication times, and so on.
4.3 Modeling of the digital resource

Tremendous enduring interest in the Dunhuang grottoes has resulted in the proliferation of many digital resources of different data types, including multimedia resources such as pictures, audio, video, and 3D models. These digital resources provide important documentation and analysis of the historical and cultural knowledge contained in the Dunhuang grottoes. It mainly includes names, types, description formats, sources, and so on.

4.4 Modeling of the mural content

The knowledge of the mural content is a representation of the content and themes depicted in the mural based on the grotto’s structure knowledge. The attributes of the mural content category include the mural content description, mural type, painting technique, and the theme expressed in the mural’s content. The theme is a semantic understanding of the content of the mural. The semantic content of the mural can, therefore, be identified as the mural’s theme.

4.5 Modeling of the historical cultural knowledge

Dunhuang grottoes related historical data and research literature constitute the main data sources of its historical and cultural knowledge. The ontology model expresses and reveals the historical and cultural knowledge of Dunhuang grottoes from different angles. It is based on people, things, times, places, and events. From the perspective of the ontology model, the historical and cultural knowledge is contained in multiple categories of rich semantic relations, and the knowledge facts and logical relationships of the historical culture are expressed from multiple angles. The facet of Dunhuang CHI mainly includes four categories: people, things, times, and places. The time category contains two sub-categories: dynasty and the epoch year. The epoch year can represent the beginning and end years of the relevant dynasty.

5.0 Knowledge graph construction of Dunhuang

The knowledge representation and organization method based on the knowledge graph should include the complete process of data collection and processing, data semantic extraction, entity association, and knowledge graph visualization and retrieval. The process of constructing the knowledge graph is shown in Figure 3. The knowledge resource layer is the foundation of the project. In this layer, structured, semi-structured, and unstructured data are collected and organized to build the original corpus based on the five previously described ontological facets. There are two main steps in the knowledge representation and organization layer. The first step is to clean the unstructured original corpus, filter out effective information resources, and annotate and extract the data resources according to the Dunhuang Grotto Ontology Model. The second step is to associate structured data, semi-structured data, and open external data through data mapping. The next layer, the knowledge service layer, consists of building a knowledge service-oriented KOS based on the completed knowledge database. This paper builds a knowledge service platform for Dunhuang CH by means of our knowledge graph. This knowledge service platform provides functions such as knowledge navigation, visual analysis, and semantic retrieval.

5.1 Data acquisition and preprocessing

The resources related to Dunhuang CH are numerous and widely available. According to the degree of structure, they can be divided into structured data, semi-structured data, and unstructured data. These data categories range from multimedia resources such as pictures and videos, Dunhuang research documents, to books and other resources. Structured data specifically includes Dunhuang related bibliographies, citation indexes, and metadata registries. Semi-structured data includes unstructured metadata descriptions and annotated resources. Finally, unstructured data is mainly multimedia resources and descriptive text of grotto structure content.

In recent years, the Dunhuang Academy has established a large number of databases, including the full-text database of the Dunhuang Conference, and the Dunhuang Studies Book Catalog Database. At the same time, through digital technology, the Dunhuang Academy has preserved a large number of high-definition images of Dunhuang murals, 3D models, and other visual resources. We were, therefore, able to obtain structured, semi-structured, and unstructured data from these data sources through cooperation with Dunhuang Academy. We also cooperated with CNKI (China National Knowledge Infrastructure)—the largest literature database in China. CNKI includes more than twenty-nine million records. Humanities, social science, natural science, and application science researchers use CNKI as a network publishing platform to supplement their academic works. This data was harvested, cleaned, and transformed to obtain valid Dunhuang CH data for this study.

5.2 Knowledge representation and organization

In order to accurately express the spatial structure of the Dunhuang grottoes according to the ontology, we extracted the description of the Dunhuang grottoes’ structural knowledge in the General Contents of Dunhuang Grot-
In this step, the unstructured text data composed of the detailed description of the structure and content of each cave was converted into structured knowledge. To assist researchers in their semantic annotation of Dunhuang domain knowledge, we developed a fine-grained text semantic annotation tool that used predefined ontologies.

The structure space of the grotto contains a lot of information. There is a strong semantic relationship between the different composition structures and the objects in the grotto. The semantic division of the spatial structure can be represented by a three-dimensional space structure, which can then be transformed into structured knowledge of spatial structures according to the spatial structure ontology (Figure 4).

Once the spatial structure of the cave has been semantically extracted, it is possible to integrate the other four facets of knowledge into this ontology, including the facets of academic works knowledge, multimedia resources knowledge, history and culture knowledge, and the facets of mural content knowledge.
When implementing the description of the semantic relationship between different entities in Dunhuang CH, it is also necessary to consider establishing association with external data as much as possible, so as to make the content of the knowledge base more abundant and accurate. In this paper, CBDB, Dunhuang Mural Thesaurus (DMT) (Wang et al. 2020b), AAT, GeoNames, and Wikipedia were selected for entity association. DMT is a thesaurus based on the Dunhuang Dictionary and other literature related to Dunhuang grottoes, which is constructed through the top-down and bottom-up methods, with AAT as the reference. The thesaurus provides a detailed division and interpretation of Dunhuang-related vocabulary.

5.3 Knowledge graph service platform

After organizing the CHI and storing it in the knowledge base, we constructed a platform for displaying our knowl-
The knowledge graph of Dunhuang grottoes. This platform provides functions such as knowledge navigation, semantic retrieval, and visualization. This platform will be opened and provided to scholars as well as members of the general public who are interested in Dunhuang CH. Under the guidance of standardized knowledge organizations and domain experts, this platform can provide standardized and accurate knowledge services for Dunhuang CH.

Within the knowledge graph, nodes represent instances, and the connection between two nodes represents the semantic relationship between entities. Through the semantic relationship between nodes, knowledge inference and mining can be achieved. The navigation bar supports retrieval operations by cave number and dynasty. For example, Figure 4 illustrates how the information linked to Cave No.61 is retrieved, as well as how the spatial structure information of the grotto is displayed in the system, allowing a clear visualization of the structural distribution of the grotto. By clicking on different entity nodes, the metadata information, research literature information, and related multimedia resources related to the node are displayed.

Semantic retrieval is an important aspect of providing knowledge services, which can help users quickly locate desired knowledge. This knowledge graph platform also supports semantic search functions. For example, if you search for “Guanyin Bodhisattva,” all caves containing the search term will be displayed on the interface, and knowledge discovery can be realized by clicking on the corresponding cave. This search is illustrated in Figure 5, steps one and two.

Because the knowledge graph is associated with external data, such as the Dunhuang Mural Thesaurus, users can click on the corresponding entity term to view externally associated data if they need to go deeper or expand on some content when using it. When you click “Bodhisattva,” you can jump to the Dunhuang fresco thesaurus to view the interpretation of the word. Likewise, when you click on a research resource, you will be redirected to the detailed page of the document in CNKI. Thus, external linkages enrich the knowledge graph and provide searchers with additional context and avenues for further exploration and clarification.

6.0 Discussion

In order to realize the knowledge organization and management of Dunhuang CH resources, this paper analyzes the characteristics of Dunhuang CH and proposes a knowledge ontology model for Dunhuang grottoes. The Dunhuang Grotto Knowledge Graph was constructed to realize the semantic representation and association of Dunhuang grottoes CHI and can provide high-quality knowledge services with distinct advantages. The following section summarizes the ideas presented in this study and discusses how they
could be further expanded and optimized by future projects.

1) In CH, building structure information is an indispensable part of architecture, but the current research on the organization of Dunhuang CHI has mainly focused on Dunhuang’s murals and manuscripts while ignoring the rich semantic information of the cave structure itself. Therefore, a spatial structure representation ontology based on the semantic features of Dunhuang grottoes was constructed in this paper.

2) When describing the semantic contents of CH, the composition of the resources can be clearly recognized by dividing the knowledge of different topics or types, which provides a reference for accurate and reasonable ontology construction. For example, this article describes the contents of Dunhuang CH through five knowledge facets, which provided guidance for the subsequent construction of a Dunhuang Grotto Knowledge Graph.

3) The content of the murals in the Dunhuang grottoes is rich in semantic information. This information needs to be represented and mined through a more refined semantic annotation model. The focus of this paper is to organize and correlate the multi-source heterogeneous and multimodal resources of Dunhuang CH. The representation of mural contents was performed at a medium granularity level. The system could be integrated with a digital interactive display system in the future to facilitate deeper understanding of the plot, action, and context of mural content.

4) The evaluation of the knowledge graph is an indispensable step of its construction. In this paper, to ensure the accuracy, completeness, and usability of the knowledge graph, we invited experts in the field of Dunhuang CH to conduct a preliminary assessment of the Dunhuang CH knowledge graph. In order to expand the scale of the knowledge graph, enhance the degree of relevance of the knowledge in the graph, and provide richer and more personalized functions, further systematic evaluation of the knowledge graph will be carried out through questionnaire surveys, user interview and observation, and eye-tracking experiments. The results of these evaluations will be examined in future research.

7.0 Conclusion

This paper analyzes the current state of Dunhuang DHI from the organization of CHI, representation of building structure, to existing CH knowledge services. A Dunhuang Grotto Ontology (covering cave structure, academic work, multimedia resources, mural content, and historical and cultural) was also designed to address current challenges in Dunhuang CHI. A multi-source, heterogeneous, multi-modal Dunhuang Grotto Knowledge Graph was subsequently constructed for the purpose of representation and association. In addition, a knowledge graph platform was built to provide knowledge service.

In this paper, the Dunhuang Grotto Knowledge Graph was realized via constructing a high-level conceptual ontology, semantic annotation, extracting entities and relationships of Dunhuang CHI, linking the knowledge of five facets, and storing them into Neo4j. By constructing a visualization and retrieval service platform, this study may promote the research, teaching, and dissemination of Dunhuang CH and provide a methodological reference for related CHI projects. In the future, a larger dataset and richer knowledge network system of Dunhuang CHI will be constructed. Varied data resources will be fused and reconstructed into smart big data to meet the demand for multiple knowledge services under different retrieval tasks and usage scenarios.

References


Books Recently Published

Compiled by J. Bradford Young

DOI:10.5771/0943-7444-2020-7-616


Publisher
Ergon – ein Verlag in der Nomos Verlagsgesellschaft mbH
Waldseestraße 3-5
D-76530 Baden-Baden
Tel. +49 (0)7221-21 04-667
Fax +49 (0)7221-21 04-27
Sparkasse Baden-Baden Giagtrag
IBAN: DE05 6625 0030 0005 0022 66
BIC: SOLADES1BAD

KNOWLEDGE ORGANIZATION
Journal of the International Society for Knowledge Organization
Richard P. Smiraglia, Editor-in-Chief
K0editor-in-chief@knoworg.org

Instructions for Authors
Manuscripts should be submitted electronically (in Microsoft® Word format) in English only via ScholarOne at https://mc04.manuscriptcentral.com/jisko. Manuscripts that do not adhere to these guidelines will be returned to the author for resubmission in proper form. Manuscripts should be accompanied by an indicative abstract of approximately 250 words. Manuscripts of articles should fall within the range 6,000–10,000 words. Longer manuscripts will be considered on consultation with the Editor-in-chief. A separate title page should include the article title and the author’s name, postal address, and E-mail address. Only the title of the article should appear on the first page of the text. Contact information must be present for all authors of a manuscript.
To protect anonymity, the author’s name should not appear on the manuscript. Criteria for acceptance will be appropriateness to the field of knowledge organization (see “Scope” at https://www.isko.org/ko.html), taking into account the merit of the contents and presentation. It is expected that all successful manuscripts will be well-situated in the domain of knowledge organization, and will cite all relevant literature from within the domain. Authors are encouraged to use the KO literature database at http://www.isko.org/lit.html.
The manuscript should be concise and should conform to professional standards of English usage and grammar. Authors whose native language is not English are encouraged to make use of professional academic English-language proofreading services. We recommend Vulpine Academic Services (vulpineacademic@gmail.com).
Manuscripts are reviewed with the understanding that they have not been previously published, are not being submitted for publication elsewhere, and that if the work received official sponsorship, it has been duly released for publication. Submissions are refereed, and authors usually will be notified within 6 to 8 weeks.

Manuscript Formatting
Under no circumstances should the author attempt to mimic the presentation of text as it appears in our published journal. Instead, please follow these instructions:
In Microsoft® Word please set the language preference (“Tools,” “Language”) to “English (US)” or “English (UK).”
The entire manuscript should be double-spaced, including notes and references.
The text should be structured with decimally-numbered subheadings (1.0, 1.1, 2.0, 2.1, 2.1.1, etc.). It should contain an introduction, giving an overview and stating the purpose, a main body, describing in sufficient detail the materials or methods used and the results or systems developed, and a conclusion or summary.
Author-generated keywords are not permitted.
Footnotes are not allowed but endnotes may be used. Do not use automatic footnotes formatting. Instead, insert a superscript numeral (Format, Font, Superscript) and create the text of the note manually.
Notes are permitted only for phrases from languages other than English, and for the titles of published works.
Bold type is not permitted.
Em-dashes should not be used as substitutes for commas. Dashes must be inserted manually (Insert, Special Characters, Em-dash) with no spaces on either side. Illustrations should be embedded within the document. Photographs (including color and halftones) should be scanned with a minimum resolution of 600 dpi and saved as .jpg files. Tables should be identified parenthetically with their URLs but not with references unless a specific page of a specific website is being quoted.

References
Reference citations within the text should have the form: (Author year). For example, (Jones 1990).
Specific page numbers are required for quoted material, e.g. (Jones 1990, 100). A citation with two authors would read (Jones and Smith 1990); three or more authors would be (Jones et al. 1990). When the author is mentioned in the text (e.g., only the initials and optional page number should appear in parentheses: According to Jones (1990), ... or “Smith et al. (2014), …). A subsequent page reference to the same cited work (e.g., to Smith 2010) should have the form “(229).” There is never a comma before the date.
In text citations should not be routinely placed at the end of a sentence or after a quotation, but an attempt should be made to work them into the narrative. For example: Jones (2010, 114) reported statistically significant results.

Footnotes are not allowed but endnotes may be used. Do not use automatic footnote formatting.

Examples:

KO is published by Ergon.

Annual subscription 2020:
Print online + (8 issues/annum; unlimited access for your Campus via Nomos eLibrary) € 159.00 incl.
Prices do not include postage and packing
Cancellation policy: Termination within 3 months’ notice to the end of the calendar year
KO KNOWLEDGE ORGANIZATION

Official Journal of the International Society for Knowledge Organization
International Journal devoted to Concept Theory, Classification, Indexing and Knowledge Representation

Scope

The more scientific data is generated in the impetuous present times, the more ordering energy needs to be expended to control these data in a retrievable fashion. With the abundance of knowledge now available the questions of new solutions to the ordering problem and thus of improved classification systems, methods and procedures have acquired unforeseen significance. For many years now they have been the focus of interest of information scientists the world over.

Until recently, the special literature relevant to classification was published in piecemeal fashion, scattered over the numerous technical journals serving the experts of the various fields such as:

- philosophy and science of science
- science policy and science organization
- mathematics, statistics and computer science
- library and information science
- archivistics and museology
- journalism and communication science
- industrial products and commodity science
- terminology, lexicography and linguistics

Beginning in 1974, KNOWLEDGE ORGANIZATION (formerly INTERNATIONAL CLASSIFICATION) has been serving as a common platform for the discussion of both theoretical background questions and practical application problems in many areas of concern. In each issue experts from many countries comment on questions of an adequate structuring and construction of ordering systems and on the problems of their use in opening the information contents of new literature, of data collections and survey, of tabular works and of other objects of scientific interest. Their contributions have been concerned with

1. clarifying the theoretical foundations (general ordering theory/science, theoretical bases of classification, data analysis and reduction)
2. describing practical operations connected with indexing/classification, as well as applications of classification systems and thesauri, manual and machine indexing
3. tracing the history of classification knowledge and methodology
4. discussing questions of education and training in classification
5. concerning themselves with the problems of terminology in general and with respect to special fields.

Aims

Thus, KNOWLEDGE ORGANIZATION is a forum for all those interested in the organization of knowledge on a universal or a domain-specific scale, using concept-analytical or concept-synthetic approaches, as well as quantitative and qualitative methodologies. KNOWLEDGE ORGANIZATION also addresses the intellectual and automatic compilation and use of classification systems and thesauri in all fields of knowledge, with special attention being given to the problems of terminology.

KNOWLEDGE ORGANIZATION publishes original articles, reports on conferences and similar communications, as well as book reviews, letters to the editor, and an extensive annotated bibliography of recent classification and indexing literature.

KNOWLEDGE ORGANIZATION should therefore be available at every university and research library of every country, at every information center, at colleges and schools of library and information science, in the hands of everybody interested in the fields mentioned above and thus also at every office for updating information on any topic related to the problems of order in our information-flooded times.

KNOWLEDGE ORGANIZATION was founded in 1973 by an international group of scholars with a consulting board of editors representing the world’s regions, the special classification fields, and the subject areas involved. From 1974-1980 it was published by K.G. Saur Verlag, München. Back issues of 1978-1992 are available from ERGON-Verlag, too.

As of 1989, KNOWLEDGE ORGANIZATION has become the official organ of the INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION (ISKO) and is included for every ISKO-member, personal or institutional in the membership fee.

Annual subscription 2020: Print + online (8 issues/ann.; unlimited access for your Campus via Nomos eLibrary) € 359,00/ann. Prices do not include postage and packing. Cancellation policy: Termination within 3 months’ notice to the end of the calendar year.

Ergon – ein Verlag in der Nomos Verlagsgesellschaft mbH, Waldseestraße 3-5, D-76530 Baden-Baden, Tel. +49 (0)7221-21 04-667, Fax +49 (0)7221-21 04-27, Sparkasse Baden-Baden Gaggenau, IBAN: DE05 6625 0030 0005 0022 66, BIC: SOLADES1BAD

Founded under the title International Classification in 1974 by Dr. Ingebrit Dahlberg, the founding president of ISKO. Dr. Dahlberg served as the journal’s editor from 1974 to 1997, and as its publisher (Indeks Verlag of Frankfurt) from 1981 to 1997.

The contents of the journal are indexed and abstracted in Social Sciences Citation Index, Web of Science, Information Science Abstracts, INSPEC Library and Information Science Abstracts (LISA), Library, Information Science & Technology Abstracts (EBSCO), Library Literature and Information Science (Wilson), PASCAL, Referativnyi Zhurnal Informatika, and Sociological Abstracts.