Jian Qin
School of Information Studies, Syracuse University, 4-187 Center for Sci & Tech, Syracuse, New York 13244, USA   Email: jqin@syr.edu

Evolving Paradigms of Knowledge Representation and Organization: A Comparative Study of Classification, XML/DTD, and Ontology

Abstract

The different points of views on knowledge representation and organization from various research communities reflect underlying philosophies and paradigms in these communities. This paper reviews differences and relations in knowledge representation and organization and generalizes four paradigms—integrative and disintegrative pragmatism and integrative and disintegrative epistemologism. Examples such as classification, XML schemas, and ontologies are compared based on how they specify concepts, build data models, and encode knowledge organization structures.

1. Introduction

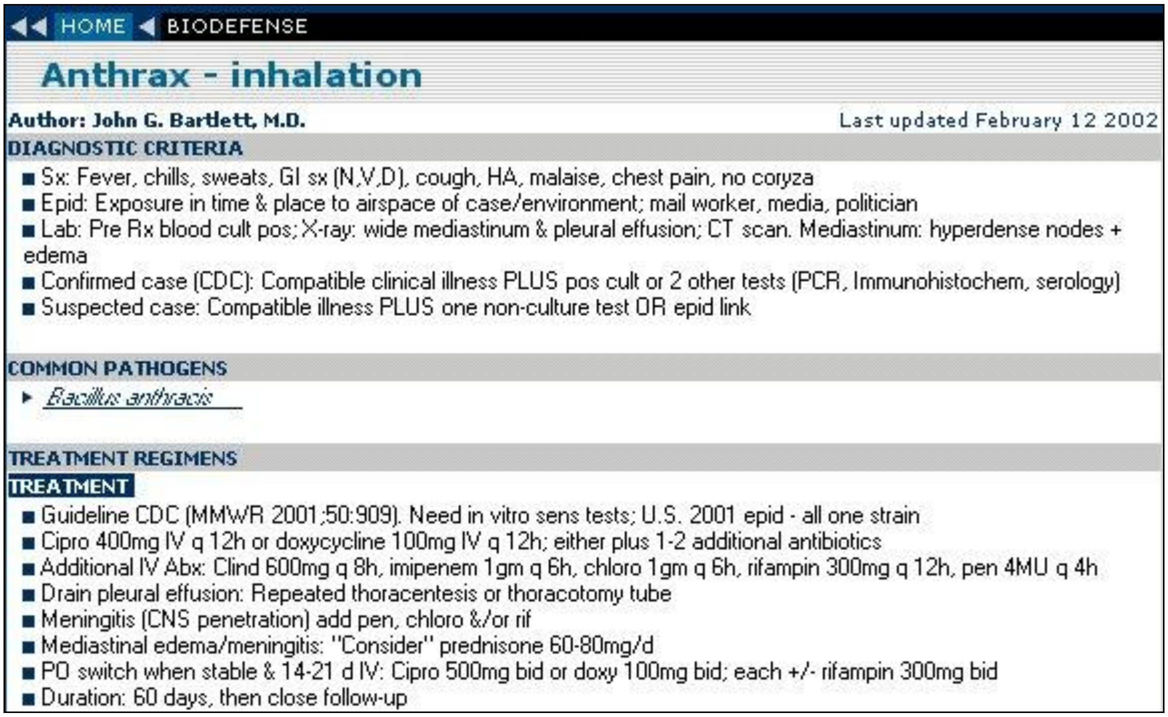
Knowledge representation (KR) is a term that several research communities use to refer to somewhat different aspects of the same research area. The artificial intelligence (AI) community considers KR as simply “something to do with writing down, in some language or communications medium, descriptions or pictures that correspond in some salient way to the world or a state of the world” (Duce & Ringland, 1988, 3). It emphasizes the ways in which knowledge can be encoded in a computer program (Bench-Capon, 1990). For the library and information science (LIS) community, KR is literally the synonym of knowledge organization, i.e., KR is referred to as the process of organizing knowledge into classifications, thesauri, or subject heading lists. KR has another meaning in LIS: it “encompasses every type and method of indexing, abstracting, cataloguing, classification, records management, bibliography and the creation of textual or bibliographic databases for information retrieval” (Anderson, 1996, 336). Adding the social dimension to knowledge organization, Hjørland (1997) states that knowledge is a part of human activities and tied to the division of labor in society, which should be the primary organization of knowledge. Knowledge organization in LIS is secondary or derived, because knowledge is organized in learned institutions and publications. These different points of views on KR suggest that an essential difference in the understanding of KR between both AI and LIS lies in the source of representation—whether KR targets human activities or *derivatives* (knowledge produced) from human activities. This difference also decides their difference in purpose—in AI KR is mainly computer-application oriented or pragmatic and the result of representation is used to support decisions on human activities, while in LIS KR is conceptually oriented or abstract and the result of representation is used for access to derivatives from human activities.

Despite the essential difference, these different versions of KR share some common principles and methodologies. For example, AI’s KR stresses adequacy and expressiveness, e.g., a scheme that represents a knowledge domain should be sufficient to allow any fact of interest to be inferred. Similarly, LIS's KR emphasizes the importance of representing the same phenomenon in different contexts such as in sociology, economics, psychology, history, and so forth. Both use some encoding language and format for representation. This paper discusses KR in such a general way as described by Duce and Ringland (1988), that is, from a structural and language point of view rather than a computational point of view. By using examples of various knowledge structures, this paper presents four paradigms prevailing in KR research and practices and compares three knowledge organization structures to demonstrate how these paradigms impact them.

2. Paradigms in Knowledge Representation and Organization

Paradigms symbolize meta-theoretical assumptions about the nature of the subject of study (Berrell & Morgan, 1979), or "universally recognized achievements that for a time provide model problems and solutions to a community of practitioners" (Kuhn, 1970). Hirschheim and Klein (1989) resolve the differences between Berrell and Morgan and Kuhn by pointing out that a paradigm consists of a "most fundamental set of assumptions adopted by a professional community that allows its members to share similar perceptions and engage in commonly shared practices." Even though differences exist in KR practices, common approaches are used across research communities. These common practices include hierarchical organization of concepts and horizontal relations between them. Let us examine the following examples of representing the concept of anthrax.

Example 1. Anthrax by diagnosis from the *Antibiotic Guide* (http://www.hopkins-abxguide.org/)



Anthrax - inhalation

Author: John G. Bartlett, M.D.
Last updated February 12 2002

DIAGNOSTIC CRITERIA

- Sx: Fever, chills, sweats, GI sx (N,V,D), cough, HA, malaise, chest pain, no coryza
- Epid: Exposure in time & place to airspace of case/environment; mail worker, media, politician
- Lab: Pre Rx blood cult pos; X-ray: wide mediastinum & pleural effusion; CT scan. Mediastinum: hyperdense nodes + edema
- Confirmed case (CDC): Compatible clinical illness PLUS pos cult or 2 other tests (PCR, Immunohistochem, serology)
- Suspected case: Compatible illness PLUS one non-culture test OR epid link

COMMON PATHOGENS

- *Bacillus anthracis*

TREATMENT REGIMENS

TREATMENT

- Guideline CDC (MMWR 2001;50:909). Need in vitro sens tests; U.S. 2001 epid - all one strain
- Cipro 400mg IV q 12h or doxycycline 100mg IV q 12h; either plus 1-2 additional antibiotics
- Additional IV Abx: Clind 600mg q 8h, imipenem 1gm q 6h, chloro 1gm q 6h, rifampin 300mg q 12h, pen 4MU q 4h
- Drain pleural effusion: Repeated thoracentesis or thoracotomy tube
- Meningitis (CNS penetration) add pen, chloro &/or rif
- Mediastinal edema/meningitis: "Consider" prednisone 60-80mg/d
- PO switch when stable & 14-21 d IV: Cipro 500mg bid or doxy 100mg bid; each +/- rifampin 300mg bid
- Duration: 60 days, then close follow-up

The *Antibiotic Guide* represents the concept of anthrax using a problem-solving approach. It divides issues surrounding anthrax into problem-solving areas such as diagnostic criteria, common pathogens, treatment regimens, and important points. Under each problem-solving area, more specific concepts and solutions are defined. In MeSH, the concept of anthrax is represented through two tree structures: the *Bacteria* and the *Bacterial Infections and Mycoses*. Each of them is a typical hierarchy that integrates into a system with different levels of knowledge about anthrax.

Example 2. The Concept of Anthrax in the *Medical Subject Headings (MeSH)*

Bacteria	Bacterial Infections and Mycoses
Endospore-Forming Bacteria	Bacterial Infections
Gram-Positive Endospore-Forming Bacteria	Gram-Positive Bacterial Infections
Gram-Positive Endospore-Forming Rods	Bacillaceae Infections
Bacillaceae	Anthrax
Bacillus	
Bacillus anthracis	

These two examples pose an important and interesting question: Are they fundamentally different in representing the knowledge or are the representations simply some variations of the same paradigm that originates from the same principles or philosophies? Obviously, the answer to this question may not be a straightforward "yes" or "no." Organizing knowledge in libraries has a long history of using an integrated approach. Think about hierarchical and faceted classifications. Both structures integrate human knowledge into a systematic arrangement, in which concepts and structures tend to be abstract and have an epistemological orientation. On the contrary, newer knowledge technologies such as XML schemas and ontologies take an opposite approach in representing knowledge, which disintegrate parts of knowledge into a problem-solving focused structure and are more pragmatic and application-oriented.

If we put these approaches together with two intercepting spectra, we obtain four paradigms as shown in Figure 1. The integration paradigm is best summarized in the theory of "integrative levels" (Feibleman, 1954). The integrative levels, as Feibleman states, represent some uniformity in science as well as the physical world. The integrative levels theory views each level of the physical world as an organization of the level or levels below it plus one emergent quality. The integrative levels are cumulative upward and complexity of the levels also increases upward; the higher level depends upon the lower, and the lower is directed by the higher. For an organization at any given level, its mechanism lies at the level below and its purpose at the level above. The MeSH example is a good demonstration of the integrative levels theory.

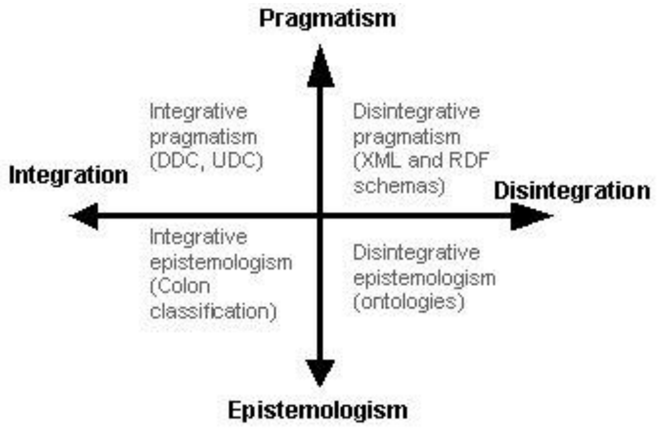

Pragmatism
Integrative pragmatism (DDC, UDC)
Disintegrative pragmatism (XML and RDF schemas)
Integration
Disintegration
Integrative epistemologism (Colon classification)
Disintegrative epistemologism (ontologies)
Epistemologism

Figure 1. Paradigms in knowledge representation and organization

The disintegration paradigm takes an opposite approach. Rather than organizing knowledge into a vertical hierarchy, disintegrative representations focus on the concept and all aspects related to it, which Ingwersen calls it "*polyrepresentation*" (1994). Disintegrative representations define a concept and match the right solutions to problems related to this concept. Consequently, the representations become the conceptual model or framework for an application. In this situation, where the concept is located in the knowledge system is less important than what solution areas there are in relation to the concept. The second example above demonstrates such an underlying statement.

Another way of considering KR paradigms is as pragmatic versus epistemological. "Pragmatism stresses the instrumentality of human knowledge and concepts." (Hjørland, 1997, 97) It takes practical consequences as the criteria of knowledge and meaning. Epistemologism engages in abstract and epistemological representations and structures. The intercepting areas as shown in Figure 1 form four distinctive paradigms*: integrative pragmatism* for which Dewey Decimal Classification (DDC) and Universal Decimal Classification are representative; *integrative epistemologism* as reflected in Colon Classification; *disintegrative pragmatism* showing a trend in newer knowledge technologies; and *disintegrative epistemologism* as represented by ontologies. Due to space limitations, an in-depth discussion of these intercepting paradigms will have to be given in another paper. They nevertheless raise a number of important questions: Do the paradigms underlie practices in both LIS and AI communities? In what ways the two communities perceive the paradigms? How have the paradigms affected the KR research and practices? Although addressing these questions is beyond the scope of this paper, a comparison between some KR structures is provided below to show how different paradigms might have influenced the representation outcomes.

3. Comparison of Knowledge Structures

Table 1 lists similarities and dissimilarities between three knowledge structures organized by the ways in which concepts are specified, data modeled, and knowledge is encoded. The variations among them are largely decided by the purpose of each representation. Classification is commonly used in almost every field of human activities and the physical world, including newer representation structures such as XML Document Type Definition (DTD) or schemas and ontologies. Library classification such as DDC (including some thesauri that have a covert classification hierarchy through a broader term/narrower term network) primarily uses a hierarchical structure to represent knowledge. Dewey's intention was to create a practical tool for librarians for matching the subject content of publications to the classification structure. Thus library classification is more concerned with how concepts are structured in order to group like materials together for easy browsing and retrieval. This means that while library classification takes an integrated approach, it is also practical. A library classification usually is not concerned with whether a concept is covered in a library's collection, but more with whether or not the knowledge structure covers all components at each level and all their aspects.

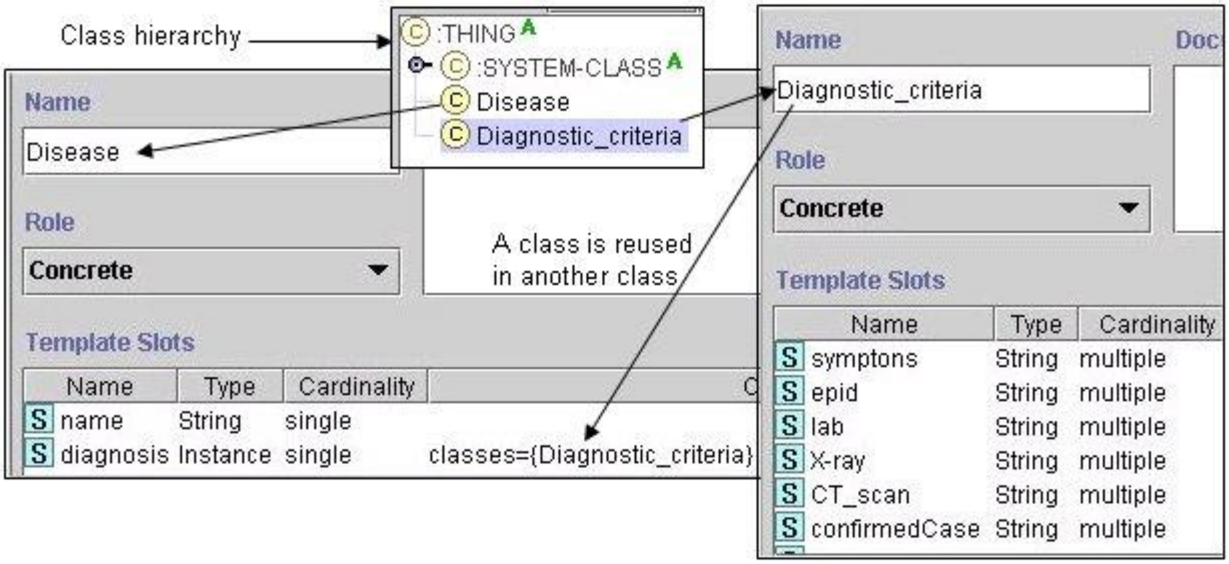
Similar to classification, XML schemas organize concepts into a hierarchy, but they are more data-oriented. That is, XML schemas show a very strong tendency for representing concepts involved in an application domain and view these concepts at the logical level. This mandates that an XML schema must provide a conceptual model for a domain by specifying what concepts there are, what the attributes are for each concept, and in which way concepts are related in an application domain.

The ways ontologies specify concepts are similar to those of XML schemas in that they are both application oriented. Because of this, ontologies in general are not intended for representing the complete human knowledge system but instead, the concepts useful to an application system. A unique method used in ontologies for creating relations between concepts is using complex slot types such as class and instance (Figure 2). Such complex slot types provide deeper representation for the multi-dimensions of concepts. Compared to the other two structures in Table 1, ontologies provide a fuller range of mechanisms for representing and organizing knowledge.

Table 1. Comparison of library classification, XML DTD/Schema, and ontology

Feature	**Library classification**	**XML DTD/Schema**	**Ontology**
Purpose	Knowledge structure for organizing library materials	Data model for organizing data	Conceptual model for a knowledge and/or application domain
Concept specification			
Structure	Hierarchical	Hierarchical	Hierarchical
Concept labeling	Class name	Element name	Class name
Concept definition	Scope note	Comment	Documentation
Concept attributes	Subdivided-by criteria or facets	Element attributes	Class slots
Attribute type	N/A	Text-based	Base data types and complex data types
Relations between concepts	See, see also	Entity, ID, IDREF	Inheritance, slot type for class and instance, inclusion of other ontologies
Data modeling			
Data structure	N/A	Relational, Object-Oriented	Relational, Object-Oriented
Data type	N/A	Character string	SQL compliant, non-SQL data types
Representation language			
Definition language	Natural language	Natural language and/or controlled vocabulary	Natural language and/or controlled vocabulary
Markup language	N/A	XML	RDF(S), DAML+OIL
Mathematic language	N/A	N/A	First-order logic

Note: RDF(S) = Resource Description Framework (Schema); DAML = DARPAR Agent Markup Language; OIL = Ontology Interchange Layer.



Name: Disease

Role: Concrete

Template Slots

Name	Type	Cardinality	C
S name	String	single	
S diagnosis	Instance	single	classes={Diagnostic_criteria}

Name: Diagnostic_criteria

Doc

Role: Concrete

Template Slots

Name	Type	Cardinality
S symptons	String	multiple
S epid	String	multiple
S lab	String	multiple
S X-ray	String	multiple
S CT_scan	String	multiple
S confirmedCase	String	multiple

Figure 2. An ontology example: a class is used as the value domain for another class

4. Discussion and Conclusions

Classification, XML schemas, and ontologies share some common approaches to representing and organizing knowledge, but they are produced under different paradigms and serve different purposes. The comparison of these knowledge structures indicates that more recent approaches to knowledge representation and organization are developed using the foundations established by precursors. Classification existed long before the computer was invented. While classifications are still being used and developed, technological advances motivated the evolution of newer knowledge- representation paradigms, which in turn generated new structures, such as XML schemas and ontologies. As shown on the paradigm chart (Figure 1), development in one area may not always move along a single direction as indicated by the direction of the arrows. Any paradigm can also move inward towards the center.

This analysis of KR paradigms is only preliminary. Before we can fully describe the model suggested here, the questions raised need to be studied not only in the context of the examples used in this paper, but also in extended examples from other domains.

References

Anderson, J.D. (1996). Organization of knowledge. In: J. Feather & P. Sturges (Eds.), *International Encyclopedia of Library and Information Science*, 336-352. London & New York: Routledge.

Bench-Capon, T.J.M. (1990). *Knowledge Representation: An Approach to Artificial Intelligence*. London: Academic Press.

Berrell, G. & Morgan, G. (1979). *Sociological Paradigms and Organizational Analysis*. London: Heinemann.

Duce, D. & Ringland, G. (1988). “Background and introduction.” In: G.A. Ringland and D.A. Duce (Eds.), *Approaches to Knowledge Representation: An Introduction*, 1-12. New York: John Wiley and Sons.

Feibleman, J. K. (1954). Theory of integrative levels. British Journal for the Philosophy of Science, 5: 59-66. (Reprinted in: L. M. Chan, Richmond, P. A., and Svenonius, E. (Eds.), *Theory of Subject Analysis: A Sourcebook*, 138-143. Littleton, Colo., Libraries Unlimited, 1989.)

Hjørland, B. (1997). *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Westport, Connecticut & London: Greenwood Press.

Hirschheim, R. and Klein, H. K. (1989). Four paradigms of information systems development. *Communications of the ACM*, 32: 1199-1216.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In: *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development of Information Retrieval*, 101-110. New York: Spring-Verlag.

Kuhn, T. (1970). *The Structure of Scientific Revolutions*. 2nd ed. Chicago: University of Chicago Press.