Tennis, J. T. and Jacob, E. K. "Toward a Theory of Structure in Information Organization Frameworks." (2008). In *Culture and Identity in Knowledge Organization: Proceedings of the 10th International Conference for Knowledge Organization.* (Montreal, Quebec August 5-8, 2008). Advances in Knowledge Organization vol. 11. Ergon: Würzburg: 262-268.

Joseph T. Tennis
Elin K. Jacob

Toward a Theory of Structure in Information Organization Frameworks

Abstract

This paper outlines a formal and systematic approach to explication of the role of structure in information organization. It presents a preliminary set of constructs that are useful for understanding the similarities and differences that obtain across information organization systems. This work seeks to provide necessary groundwork for development of a *theory of structure* that can serve as a lens through which to observe patterns across systems of information organization.

Introduction

In the wake of the Semantic Web initiative, the rapid evolution of social computing and the growing interest in Second Life, we find ourselves in the midst of what might well be described as the Cambrian age of Information Science. A wealth of new technologies and innovative work practices has generated widespread interest in the development of novel approaches to organizing information, each of which is manifested within a particular context, for a specific group of individuals, in order to address a more or less explicit set of goals and objectives. This focus on the design and implementation of knowledge organization has produced systems that run the gamut from exquisitely crafted, multi-million dollar creations such as the National Cancer Institute's cancer bioinformatics grid (caBIG) (Saltz et al., 2006), and the National Library of Medicine's MEDLINE (NLM, 2007) to an emerging smorgasbord of socially-organized information systems such as Del.icio.us (del.icio.us, 2008) and Connotea (Nature, 2008). The upshot is that we are currently confronted by an unprecedented explosion in the number and variety of formal and informal systems for knowledge representation and organization.

In light of the unprecedented increase of information systems, there is growing need to be able not only to evaluate "new" representational frameworks, such as folksonomies and ontologies, but also to compare these new frameworks with more traditional systems, such as thesauri, subject heading lists, and enumerative classification schemes. While all of these representational systems share the general goal of supporting access to resources, each works in different ways to effect this end. And, although a basic understanding of these systems has evolved within the knowledge organization community, there is increasing need for a theory of structure that can provide a lens through which to compare emerging manifestations; to systematically assess their similarities and differences; to rigorously identify their strengths and weaknesses; and to detect gaps in our own understandings of the utility of these tools for organization and retrieval.

Structure in the Context of Information Organization Frameworks

The structure of a social tagging system, a metadata scheme, or an indexing language must be understood within the framework in which it occurs. The *information organization framework* itself is comprised of three distinct but interrelated components: the *discourse* that establishes the goals, priorities and values of the system; the *work practices* involved in the application and maintenance of the system; and the *structure* that instantiates both the discourses underlying the framework and the work practices that make it visible. Thus, it is the work practice(s) and discourse(s) with which a system is associated that shape apprehension and understanding of its structure. More importantly, it is the discourses and work practices of a domain that determine the structure of its information organization framework. For example, ontology curation (or engineering) is an information organization framework, and the Gene Ontology (GO) is a specific instance of ontology curation. The discourses revolving around GO reflect the fact that its work practices are focused on representation of the natural (or biological) world; and the structure of GO is therefore informed by this scientific and representationalist focus and the work practices and discourses that follow from that focus.

In order to comprehend the function of structure within an information organization framework -- to appreciate structure as the product of decisions and priorities established by work practices and discourses within a given domain -- it is necessary to begin with a robust theory of structure itself.

Structure is one of those concepts, like *information*, whose intension is simply assumed. In *Metaphors we live by* (1980), Lakoff and Johnson argue that we comprehend many of the concepts we use by imbuing them with a physical structure that emerges from our day-to-day experiences (p. 59). These "structural metaphors" reflect "systematic correlations within our experience" (p. 61) and allow us to find coherence across diverse experiences (p. 82) by structuring the dimensions of one experience in terms of the dimensions of another. Even though structure-as-metaphor is central to Lakoff and Johnson's argument, they fail to define just what they mean when they speak of *structure*. In light of their discussion of experiential gestalts, it is possible to infer that they are referring to the components -- the "dimensions" -- of experience when they speak of structure (1980, p. 83).

Sewell (1992) observes that we frequently conceive of structure as "primary, hard, and immutable, like the girders of a building" (p. 2) -- a "thing" that exists independently of our own experience but yet stabilizes and gives shape to it, like the dimensions of experience to which Lakoff and Johnson (1980) refer. But structure is far more than the sum of the components that make up our experience of an entity, an event or a system. In fact, Green (2002) stresses the impossibility of separating the idea of structure from the relationships that link the components of a system. She contends that "[s]tructure and relationships are inextricably interconnected. Wherever structure exists, relationships occur between the components of the structure. Similarly, wherever relationships exist, structure emerges" (p. 73).

Bunge (2003) takes Green's argument one step further when he points out that the concept of structure cannot stand alone. Although this is implicit in Green's contention that structure and relationships co-occur, Bunge argues that the apprehension of structure is always dependent on an existing system of relationships of which structure

itself is but a property: "Structures are properties of systems: there are no structures in themselves" (p. 277). He proceeds to define structure as "the set of all the relations among [a system's] components, particularly those that hold the system together" (p. 277), and he offers examples to illustrate his argument: "the structure of a sentence is the order of the types of its constituents, such as Subject-Verb-Object in the case of 'Socrates drank hemlock'; the structure of a theory is the relation of entailment; the structure of a DNA molecule is the sequence of the nucleotides that compose it" (p. 277).

While these intellectual precedents stress the internal nature of structure as a set of components and the relationships between them, it is also important to consider the external aspects of structure and the effects of interaction among structures, work practices and discourses. In his seminal effort to develop a theory of structure for the social sciences, Sewell (1992) defines structure as a duality or interaction comprised of schemas and resources, where schemas are virtual -- the "fundamental tools of thought, … conventions, recipes, scenarios, principles of action, and habits of speech" (p. 8) -- and resources are actual -- the "manifestations and consequences" (p. 11), the "instantiations or embodiments of schemas [that] inculcate and justify the schemas" (p. 13). And, "If structures are dual in this sense, then it must be true that schemas are the effects of resources, just as resources are the effects of schemas" (p. 13) and schemas not instantiated in or supported by resources would fade from memory, "just as resources without cultural schemas to direct their use would eventually dissipate and decay" (p. 13). Sewell subsequently defines five axioms of structure that not only follow from the dual nature of structure as both schema and resource but also capture the variability, flexibility and ultimate mutability of structure:

1) The multiplicity of structures. Multiplicity ensures flexibility and versatility in that it provides for different discourses and different work practices to adopt and apply different relational models in the design of information systems (pp. 16-17).

2) The transposability of schemas. Transposability of schemas allows for different relational models to be applied across a wide array of situations or extended to accommodate the needs of novel work practices or discourses, resulting in different structural forms (pp. 17-18).

3) The unpredictability of resource accumulation. The variability arising from the intersection of multiple structures or the extension of schemas producing different structural forms ensures that structure as the consequences of implementation can not be anticipated or predicted (p. 18).

4) The polysemy of resources. Because resources are open to interpretation in different systems, every resource is potentially ambiguous in that it is capable of being re-interpreted within different schemas (pp. 18-19).

5) The intersection of structures. The intersection of structures is one by-product of the re-interpretability of resources; but schemas as well as resources can be transposed. As Sewell observes, the "intersection of structures, in fact, takes place in both the schema and the resource dimensions" (p. 19).

In light of these five axioms, Sewell proceeds to define structures as "sets of mutually sustaining schemas and resources that empower and constrain social action and that tend to be reproduced by that social action. … [S]tructures are multiple and intersecting, because schemas are transposable, and because resources are polysemic

and accumulate unpredictably" (p. 19). Sewell's definition can be extended to provide a definition of representational systems as multiple, intersecting and potentially polysemic structures comprised of mutually sustaining discourses and work practices that accumulate unpredictably because they empower and constrain representation and are themselves reproduced, transposed and extended by the act of representation.

Three Postulates of Structure

In light of the above discussion, we propose postulates of structure in information organization frameworks. We define structure as a constructed space consisting of a set of internal partitions, each of which is connected to other partitions in the set in a meaningful way, either as a linear sequence (i.e., a continuum or process) or a network of links (i.e., a web) at the lower levels or as a hierarchical or polyhierarchical organization of part-whole and/or is-a relationships at higher levels. In the context of an information organization framework, a structure is the cohesive whole or "container" created by the establishment of qualified, meaningful relationships among the components, "whether conceptual or material, natural or social, technical or semiotic" (Bunge, 2003, p. 277), which comprise the "bounded space" of the structure.

The definition of structure as it applies to knowledge organization systems requires its own boundaries. These we provide in the form of three postulates:

Postulate 1. The smallest unit of structure is the statement.

Postulate 2. Statements are collected in levels of aggregation.

Postulate 3. The most comprehensive unit of structure is the complex.

A statement is an assertion of a relationship between a resource of interest, an attribute that can be ascribed to that resource, and the value of the attribute as it applies to the subject resource. As such, a statement is a representation of a resource that conforms to the subject-predicate-object format of a simple clause. For example, in the assertion "The title of this paper is *Toward a theory of structure in information organization frameworks*", "this paper" is the subject of the representation; "has title" is the predicate that establishes a meaningful relationship between the subject and the object; and "*Toward a theory of structure in information organization frameworks*" is the value (or object) of the predicate as it applies to the subject of the assertion.

Statements are collected within compound structures that reflect not only the increasingly more sophisticated internal relationships of statements within a structure but also the relationship of the structure itself to the discourse(s) and work practice(s) of the information organization framework with which it is associated. These *levels of aggregation* consist of *records, schemes, systems,* and *complexes.* At the simplest level of aggregation, a *record* consists of all the statements that have been made about a given resource within a given work practice. For example, the following set of statements, presented in rdf/xml syntax, constitute a record for this paper:

```
<rdf:RDF
  xmlns:dc="http://dublincore.org/2008/01/14/dcterms.rdf#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   <rdf:Description
       rdf:about="http://www.dlib.org/dlib/june98/scout/06roszkowski.html">
     <dc:title>Toward a theory of structure in information organization
      frameworks</dc:title>
```

```
    <dc:creator>Joseph T. Tennis</dc:creator>
    <dc:creator>Elin K. Jacob</dc:creator>
    <dc:date>2008</dc:date>
    <dc:type>Text</dc:type>
    <dc:language>en</dc:language>
  </rdf:Description>
</rdf:RDF>
```

A s*cheme* defines the set of predicates and classes that can be used to make an assertion about resources. Thus, statements collocated as a record are legitimated by one or more schemes that establish the boundaries of structure by constraining the set of possible statements -- the set of all possible relationships that can be established for a resource using that particular scheme. For example, the set of relationships defined as properties in the Dublin Core Metadata Element Set [DCMES] Version 1.1 (DCMI, 2008) along with the Dublin Core Abstract Model (Powell, 2008), prescribes the range of statements that can be made about a resource using DCMES; while it is possible to state the creator, the title and the publisher of a digital resource using properties from the DCMES scheme, it is beyond the scope of this scheme to assert, in machine-readable format, the location of that resource's mirror site(s).

Classification and categorization schemes used to organize collections of resources, such as Library of Congress Classification [LCC] and Library of Congress Subject Headings [LCSH], generally limit statements about resources to assertions regarding intellectual content. Thus they prescribe the range of topical statements that can be made about a resource by defining the nature and scope of classes-as-statements. For example, it can be asserted that this paper is about knowledge organization. More importantly, however, such an assertion is situated within an external structure of hierarchical or polyhierarchical relationships of classes-as-statements, from which additional statements can be inferred.

A *system* is the instantiation of all records that have been generated within the context of one or more schemes. And while the nature and scope of an individual record is constrained by interaction between the resource and the applicable scheme(s), a system is the result of creating records according to said scheme(s).

It is at the level of the *complex* that interaction occurs between the non-human resources (the records, schemes, and systems), the discourse(s) of the human agents, and the work practice(s). Here the line can be blurred between the act of creating structure and the result of that creative act, the structure itself. Like a small and localized manifestation of Foucault's épistémè (Foucault, 1980 & 1994), the complex is the dermis between solid state of structure and gaseous state of discourse.

Mooers's Method of Descriptors

An example of structure can be found in Mooers's Descriptor Method. He defines a descriptor as having parts – and in so doing defines its structure, the structure of a descriptor. Here we will abstract from Mooers (2003) in order to provide an example of structure.

Our reading of Mooers finds that a descriptor has five parts: (1) idea-element, (2) verbal expression/notation, (3) definition, (4) explicit relationship with a domain, and (5) explicit relationship to descriptor method of information retrieval. There are three

intrinsic (non-separable) elements, and two extrinsic (separable) elements. These are all elements that can generate their own statements, when taken in aggregation form a descriptor record (say for AIRPLANES). However, here we can only see two explicit statements the idea element (descriptor) and the verbal expression of the descriptor – the other parts are not explicit, but derived from context.

This record is not used alone – but is placed in a scheme/system. This scheme/system could be Mooers's Zatocoding framework (1951) or some other similar framework. And in this case we see that "airplanes" is provided along with other descriptors (fuselage, wheel, wing, etc.), under what Mooers calls a "leading question," *Is a specific component or body studied?* (Mooers, 2003). At this point we can recognize the descriptor as part of a structure used in descriptor methodology.

The dual nature of scheme/system is not altogether obvious. The scheme, for our purposes, is the specification of the set of all possible statements (in Mooers's framework, all possible idea-elements, verbal expressions/notations, etc.), and the system is the actual instantiation of those statements (in Mooers's framework, the aeronautical descriptor list from which our example is excerpted, in part). Both scheme and system are necessary for a discussion of structure because many instances of structure in information organization allow for the creation of new statements. We must rely on interaction between the scheme and the system -- the intersection of structures -- to support this creative work.

Finally this scheme/system is part of what Mooers calls the descriptor method of information retrieval, which is based on a *total systems view of the use of information* (Mooers, 2003, p. 813). This view posits that a scheme/system of descriptors will serve a specific and constrained group of uses as well as a specific and constrained collection of documents. This is the aggregation level of the complex where basic structures are deployed in a wider social structuration of schemas and resources (Sewell, 1992). Key to the descriptor method complex is the well defined purpose for using the collection described by descriptors, narrowing the interest in the resources to be retrieved (Mooers, 2003, p. 813).

Beyond complex we reach the limit of structure. That limit we call discourse. Discourse here outlines the priorities of descriptor methodology, in our example. Mooers points out the empirical discursive placement of his systems: unlike other representational systems that tend to be implemented consistently and without reference to the placement of the system, "descriptor systems are created at each installation according to a methodology embodying the utmost empiricism" (Mooers, 2003, p. 815). Another facet of the discourse surrounding Mooers' work is the focus on the idea-element, rather than terms. As counter examples to the Descriptor System, Mooers points to the Uniterm System and the Thesaurus System; and it should be possible to speciate these structures by applying the same framework used to provide an anatomy of the Descriptor System.

Structure and Its Theory

Defining structure in this way offers us the descriptive power to compare classification, ontologies, folksonomies, and web directories from the smallest level of aggregation to levels approaching discourse analysis. It alerts the analyst to these levels, and it serves as a touchstone for outlining interpretations in an increasingly

diverse universe of indexing languages. To follow these two points, identifying uniformity across variations proves useful, because a theory of structure allows us to see where and how domains form the design decisions independent of structure. We can say this because we can assume knowledge of the structure of the example indexing language, and the content and form, offered by domain analysis, become the variables. Thus a structural comparison between the Descriptor System and the Uniterm System should provide the domain analyst with a level playing field to negotiate the nuances of the particular domain represented by both systems.

A knowledge organization tool is made up of many parts. It has an anatomy. Working toward a theory of structure moves us closer to systematically understanding a key component of that anatomy: the skeleton of information organization frameworks. Structure is the scaffold of the work done by knowledge organization. And we assume we know the limits of structures in controlled vocabularies (NISO, 2005). And though this is contested by some, we see more and more complexity emerge through diversification in practice and tradition. If we want to proscribe novelty of new systems and initiatives (Soergel, 1999), then we must be clear on what grounds. Evaluation of the parts and functions of information organization frameworks stands as one route to this end. Such evaluation, based on the anatomy of these frameworks, will move us closer to a comparative analysis of utility, and a clear statement of what structures provide meaningful, and not strange, machinations.

References

Bunge, Mario (2003). *Philosophical dictionary*, enlarged edition. Amherst, NY: Prometheus.

Del.icio.us. (2007). Available at: http://del.icio.us/

Foucault, Michel (1994). *The order of things; an archaeology of the human sciences*. New York: Vintage Books.

Foucault, Michel (1980). *Power/Knowledge.* New York: Pantheon.

Green, Rebecca (2002). Internally-structured conceptual models in cognitive semantics. In R. Green, C. A. Bean & S. H. Myaeng (Eds.), *The semantics of relationships: An interdisciplinary perspective* (pp. 73-89). Dordrecht: Kluwer.

Lakoff, George & Johnson, Mark (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Mooers, Calvin N. (1951). Zatocoding applied to mechanical organization of knowledge. *American Documentation 2*(1): 20-32.

Mooers, Calvin N. (2003). Descriptors. In M. Drake (Ed.), *Encyclopedia of library and information science*, 2nd ed. (pp. 813-821). New York: Marcel Dekker.

Nature (2007). Connotea. Available at: http://www.connotea.org/

NISO (2005). Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.

NLM (2007). Medline Fact Sheet, 2007. Available at: http://www.nlm.nih.gov/pubs/factsheets/medline.html

Powell, Andy et al. (2007). DCMI Abstract Model. Available: http://dublincore.org/documents/abstract-model/

Saltz, Joel, et al. (2006). caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid. *Bioinformatics, 22*(15), 1910-1916. Available at: http://bioinformatics.oxfordjournals.org/cgi/content/full/22/15/1910

Sewell, William H. (1992). A theory of structure: Duality, agency, and transformation. *American Journal of Sociology, 98*(1), 1-29.

Soergel, D. (1999). "The rise of ontologies or the reinvention of classification." Journal of the American Society for Information Science 50(12): 1119-20.