Jacob, E., & Priss, U. (1999). Application of faceted classification structures. Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, 87-106. doi: 10.7152/acro.v10i1.12483

Application of faceted classification structures in electronic knowledge resources

Elin Jacob & Uta Priss
School of Library and Information Science
Indiana University, Bloomington, Indiana, USA

Abstract

This paper outlines a theoretical framework for the use of faceted languages in the design of information systems. It identifies controlled vocabulary, collocation and fixed citation order as fundamental principles driving traditional classification schemes and analyzes their strengths and weaknesses when applied to the organization of Web-based knowledge resources. By applying these principles within a conceptual rather than a linguistic framework, it proposes an approach to indexing that is flexible and dynamic in its application across knowledge domains.

Introduction

This paper argues that analysis of the objectives and practical application of fundamental classificatory principles such as controlled vocabulary and fixed citation order can point to alternative indexing structures that are inherently dynamic in their ability to accommodate the domain languages of various discourse communities. By juxtaposing the need for flexibility in indexing the diversity of materials on the World Wide Web against the limitations imposed by conventional classificatory practices, it is possible to isolate the central factors that make traditional classification schemes less than effective in a Web-based environment. Reanalysis of conventional practices in light of continuing problems points up the need to approach these problems from a new perspective that will generate non-traditional solutions that build on the fundamental objectives of organizational structures rather than implementing practices that have proven ineffective even in traditional environments. This paper does not argue against the fundamental tenets of controlled vocabulary and citation order. Rather, it argues that new approaches to organizing conceptual space depend upon rejecting the notions of fixed citation order and the linguistic basis of a controlled vocabulary and electing, instead, to rely on facilitation of flexible citation order and the identification and definition of concepts – of events, activities, properties and attributes -- as the basis for a faceted vocabulary. Provisions for flexibility in the ordering of facets and mapping of natural language terms within a system of well-defined concepts and relationships offer alternative solutions that

ISSN: 2324-9773

incorporate those basic principles which make controlled structures effective while accommodating the needs of a heterogeneous mix of users.

Traditional classificatory approaches

The primary objective of any classification scheme is the systematic grouping of entities such that the members of each class demonstrate an essential similarity which distinguishes them as a group from the members of any other class. Within the traditional library environment, this grouping function has encouraged the view that classification schemes serve, first and foremost, to order physical objects on the shelf. Thus the notion that bibliographical classification schemes function as an intellectual tool for demonstrating relationships of similarity based upon the intellectual content of documents has been overshadowed by their potential for generating a physical ordering of items – a "mark-and-park" structure that is linear yet physically flexible, ensuring, through the aegis of relative location, hospitality in arrangement within a constantly expanding collection.

The phenomenal proliferation of digital resources available on the World Wide Web has forced information specialists to rethink the traditional enumerative model of classification. Although traditional classification schemes effectively support access to documents housed within the physical library, the Web lacks physical properties such as dimension, metrical distance and time. It is increasingly obvious that problems of access to digital documents stored within an essentially nonphysical space cannot be resolved by methods for organizing physical space. Indeed, there is growing awareness among information specialists that effective retrieval within the nonphysical and inherently nonlinear arena of conceptual space will depend, ultimately, upon representing the intellectual content of individual documents rather than groups of documents. Furthermore, the fluid and highly dynamic environment of the World Wide Web will require the development of organizational structures that utilize flexible indexing systems such as faceted classification schemes and faceted thesauri.

Document retrieval and the world wide web

Access to digital materials on the Web is generally facilitated either by enumerative classification systems such as Yahoo! or by keyword search engines supported by hypertext linking structures. Unfortunately, enumerative schemes in the electronic environment are hampered by the very problems that plague enumerative schemes in a paper environment – problems that include the excessive length and complexity of the scheme; the static nature of class arrangement within a single hierarchical structure; and the distribution of related concepts or phenomena throughout the scheme. Furthermore, the inability of the searcher to identify the relevant class or

ISSN: 2324-9773

classes without previous knowledge of the organizational structure points up the need for an external indexing structure, such as the relative index of the Dewey Decimal Classification [DDC], that can provide subject access into the classification scheme itself.

In contrast to the complexity of a subject-based enumerative scheme, the typical search engine is a simple pattern-matching system: that is, it searches document text and/or metadata in an attempt to identify a set of documents that "match" one or more of the words in the searcher's query. Because these search engines are generally limited to simple matching of alphanumeric strings, they cannot function at the conceptual level. That is, they are incapable of distinguishing between homographs (e.g., between the nouns [horse] *bit* and [drill] *bit* and the past tense of the verb *bite*); of applying a set of synonyms that refer to the same concept; or of identifying related documents on the basis of part-whole relationships. A search based on matching of strings will frequently retrieve a large document set characterized by a high percentage of references that are not relevant to the search query. The need for collocation based on concepts has led to the development of systems such as Northern Light [www.northernlight.com], which appears to enhance retrieval performance by application of automated categorization techniques to the set of documents generated in a string-matching search.

The essential nature of current hypertext systems is similarly problematic. The purported non-linearity of hypertext linking structures has encouraged the popular notion that hyperlinks can serve as ad hoc indexing structures in a diverse conceptual space such as the World Wide Web. There are those who argue that these systems of hyperlinks, in association with full-text Boolean searching, offer a simple solution to the problems encountered in indexing highly dynamic collections of documents that cross disciplinary boundaries. (Williamson, 1998. See also essays included in "Navigating Among the Disciplines: the Library and Interdisciplinary Enquiry," *Library Trends*, vol. 51(2) edited by Palmer, 1996.) The extension of Bush's (1996/1945) argument for an associative structure organizing the individual's private storehouse of knowledge to Nelson's (1994) vision of Xanadu as an international and universal archival system has encouraged some information professionals to propose that the associative nature of hypertext systems will be capable of managing the entire body of knowledge accrued across the centuries (e.g., DeBuse, 1988).

While it is true that indexing theory may contribute to the development of more effective links in a hypertext environment (Liebscher 1994), such linking structures do not constitute an indexing system, however ad hoc. Hyperlinks do not embody a consistent structural format: that is, they may be subject-to-document links; footnote-to-document links; table-of-contents-to-chapter links; or simple, sequential page-to-page links. All hyperlinks are indexical, however. That is, just as the reference of an indexical pronoun such as *I* or *me* is determined by the speaker, a hyperlink creates a simple, one-to-one relationship between the source document and a target document –

ISSN: 2324-9773

a relationship determined either by the system builder or generated dynamically by matching of alphanumeric strings.

Traditional indexing systems are designed to collocate related documents or document surrogates, but a hypertext link can only provide immediate access to a single document and not to a group of related documents. Although pop-up menus or scroll bars using hyperlinks may appear to collocate documents, the set of documents is fixed, having been pre-selected by the system developer. When a group of documents has been retrieved in response to a particular query, each link provides access to only one document in the set. Thus, while a hypertext link may support immediate access to documents, it is not an indexing structure in that it cannot collocate conceptually-related documents.

More importantly, as Girill (1991) observes, the links created between documents in a hypertext environment are not constrained by the systematic application of an indexing language. They are determined, instead, by keyword choices within local, document-based contexts. From a global perspective, this lack of consistent representation is characteristic of hypertext structures. Within a distributed hypertext environment, it is not possible to apply a controlled vocabulary. Hyperlinks, therefore, can neither facilitate collocation of related documents nor support coherence of representation across documents. Because they do not fulfill the primary objective of indexing – collocation of conceptually-related documents through the systematic application of a well-defined representational language -- hyperlink structures can neither replace nor replicate traditional indexing systems. Rather, they should be understood as simple navigational tools supporting limited movement across conceptual space.

The role of controlled vocabulary

Representation involves the establishment of systematic correspondence between two separate domains such that an entity in the target domain is "re-presented" through the medium of the modeling domain (Barsalou, 1992). In all forms of indexing, including classification, the indexing language (the modeling domain) is used to "re-present" the conceptual content of a document (the target domain). It is also the medium used by the searcher to model or “re-present” her information request. The product of this modeling process is a surrogate that stands in for, or takes the place of, the intellectual content of the document (Jacob & Shaw, 1998), on the one hand, or the information need, on the other hand. Although no one representation of a text can ever be complete (Blair, 1990), to the extent that the surrogate corresponds to the modeled object or concept in the target domain, the two can be thought of as representationally equivalent (Barsalou, 1992). Blair (1990, 1992) points out that the effectiveness of a retrieval system relies upon the nature and quality of the document representations -- the surrogates – available to the system.

In many online retrieval systems, however, an overweening emphasis on development of interfaces that will enhance interaction between the system and the searcher has overshadowed the primary role of the representational structure in supporting communication between the indexer and the searcher. Effective retrieval requires correspondence between the language the searcher uses to model an information need and the language the indexer employs to model the intellectual content of a given document. Such correspondence depends, in turn, on the development and consistent application of a well-constructed indexing language -- a controlled vocabulary that establishes an indexical, one-to-one relationship between an authorized indexing term, or descriptor, and the concept to which it points. As Foucault (1970) observes, a well-constructed language is not susceptible to the vagaries of individual experience or prejudice because the descriptors which comprise the language are inherently indexical. That is, any given concept – any entity, activity, property or attribute -- will be represented by only one authorized term and each such term will point to only one entity, activity, property or attribute. Because a natural language frequently includes more than one linguistic term for each concept, a well-constructed indexing language may incorporate a syndetic structure of *See* or *Use* references that point the user to the appropriate descriptors.

A traditional classification scheme is just such a well-constructed language that rests on four basic assumptions: 1) each class can be defined by a set of characteristics, attributes and/or properties; 2) the set of criteria that define a class are individually necessary and jointly sufficient and will be shared by all members of the class; 3) the set of classes so defined can be ordered as an inclusive hierarchy that proceeds from the general to the specific through gradation by specialty and the establishment of generic (i.e., genus-species) relationships; and 4) such a scheme constitutes a static and unitary conceptual framework that reflects an inherent universal order. The relational structure perpetuated by a well-constructed, domain-specific vocabulary not only defines the parameters of the domain, but determines its conceptual boundaries. In so doing, it embeds a domain-specific perspective – a ruling paradigm or world view -- that is reflected in the way the particular domain conceptualizes and organizes the phenomena of investigation (Jacob 1994). Thus development of a patently universal classification scheme such as DDC reflected contemporary views in its separation of reason and imagination – of masculine and feminine -- thereby generating "a canonical map [of knowledge] that reifies established perspectives—a hegemonic discourse … reflecting the authority of mainstream experts, who, like Bacon, contrast masculine science to feminine emotion" (Olson 1994, 305).

Blair (1990) points up the domain-specific nature of language by likening words to tools. Drawing heavily on the later philosophy of Wittgenstein, he argues that words-as-tools are used in "language games" to accomplish specific tasks. Each "language game" is associated with a "form of life" – one of the "everyday human activities which make up our lives in a social sense" (Blair, 1990, 148). To play the language game requires not only familiarity with a given activity but also knowledge of how to

ISSN: 2324-9773

use the appropriate words-as-tools within the particular domain. Thus the form of life with which a particular language game is associated provides constraints that determine how words are to be used – how they "mean".

The meaning of a term is both learned from and shaped by participation in a particular language game and can be fully comprehended only through its use within a particular domain or form of life. It follows, then, that, within the conceptual framework of an indexing language, there can be no neutral terms – no terms whose meaning is determined independently of the immediate domain of activity or of its relationships with other terms in that domain. While competing terms may refer to the same concept across domains, each term will reflect the world view of the particular domain or form of life in which the term is used. Thus, the meaning of any one term can only be comprehended within the specific context of the domain – within the conceptual framework established by the indexing language. In order to organize and represent conceptual content according to the world view or ruling paradigm of the particular domain, any well-constructed indexing language is therefore constrained by the need to provide for consistency of representation through the establishment of indexicality -- a one-for-one referential structure in which each term is tied to a specific concept and assures that each concept will be represented by only one term.

Because indexicality is generally determined within a specific knowledge domain, it may not be shared by competing domains since each such "form of life" strives, through the aegis of its language game, to establish control over its particular arena of specialization. The *Diagnostic and Statistical Manual of Mental Disorders, fourth edition* (American Psychiatric Association, 1994), or DSM-IV, illustrates the central role of a well-constructed language in establishing the claims of one discipline [psychiatry] as the dominant knowledge domain within an otherwise shared arena of interest [mental disorders]. Through a systematic process of identification and validation, the developers of DSM-IV sought to embed within the domain language of psychiatry certain definitional constraints that would provide for consistent diagnoses within and across disciplinary and professional boundaries (McCarthy & Gerring, 1994). By establishing the language of psychiatry as *the* classificatory language for the entire community of mental health clinicians and academicians, the American Psychiatric Association [APA] sought to solidify psychiatry's pivotal role in the arena of mental health and to extend its scope of influence beyond its immediate disciplinary boundaries so as to encompass all clinicians and researchers in the area of mental health, regardless of disciplinary or professional affiliation. By establishing a scientific, empirically-grounded basis for its classificatory language, psychiatry sought to enhance its prestige within the broader arena of biomedicine. The former endeavor is supported by wide-spread adoption of the DSM-IV as the classificatory structure for the mental health care industry. As Jacob and Albrechtsen (1997) point out, however, the latter endeavor can only be achieved through the development of a well-constructed language that establishes indexicality through the

ISSN: 2324-9773

identification and standardization of a definitional criteria set for each and every mental disorder included in the DSM-IV.

A domain-specific vocabulary tends to preclude active understanding across domains precisely because it demands comprehension of the referential structure embedded within a rigidly defined and predetermined conceptual organization. The inability to establish indexicality across domains inhibits effective communication. But communication within the knowledge domain itself -- communication among active participants in the domain discourse -- may also be affected by the individual's domain expertise and/or familiarity with the domain vocabulary. Because the ability (or lack thereof) of certain individuals and/or discourse groups to communicate effectively with members of a knowledge domain who are more fluent in the domain language, a well-constructed language serves not only to organize the subject matter but also to restrict access to specific bodies of knowledge based on domain expertise.

Obviously, a well-constructed language – a tightly-controlled and well-defined domain vocabulary – is both an asset and a liability in any indexing system. While it serves to promote communication through the establishment of an indexical relationship between a term and its meaning, thereby supporting the objectives of collocation, it does so by demanding conformity – by limiting what can be known to that which can be represented by the set of authorized terms. In so doing, it stands as a gatekeeper, effectively limiting or denying access for those without requisite knowledge or experience of the predominant language game. A controlled vocabulary is a necessary evil, providing for correspondence between the target domain and the modeling domain while simultaneously enforcing a rigid, static and contrived environment unresponsive to the dynamism and heterogeneity that characterizes both human knowledge and natural language.

Organizing web-based collections

A universal classification such as the Dewey Decimal Classification [DDC], the Library of Congress Classification [LCC] or that created by Web-based Yahoo! constitutes a controlled vocabulary. But the rigidity of these enumerative structures renders them less than effective in the diverse and multidisciplinary environment of the Web. Traditional schemes assume a single universal reality that can be represented as a hierarchical tree structure whose classes and relationships between classes are not only predefined but also static. Olson (1994, 1996) and Frohmann (1994) have argued persuasively that the assumption of universal order inherent in such schemes – the assumption of an external and objective reality that exists independently of the classification scheme itself -- is untenable. Williamson (1998) concurs, observing that, unlike the static stability that characterizes enumerative schemes, the storehouse of human knowledge is fluid and dynamic, continuously generating new disciplines as well as new knowledge and new subjects. She points

ISSN: 2324-9773

out that the interdisciplinary nature of this expanding body of knowledge "is characterized by instability, lack of predictability, and spontaneous response to politically, socially and environmentally based issues" (118). Because traditional approaches to classification prescribe that a particular entity must be slotted in a single class within a predefined and static conceptual order, they cannot easily respond to the highly dynamic nature of human knowledge. Furthermore, when the hierarchical structure of a classification scheme is organized by discipline, aspects of a subject are scattered throughout the class structure; when the structure is organized by phenomena, however, the disciplines themselves are scattered (Williamson 1998). Organizing an unstable and dynamic environment such as the Web will require a representational scheme that can accommodate the fluidity of human knowledge and provide for flexible reconfiguration of the classificatory structure in response to a multiplicity of domain paradigms.

Postcoordinate indexing systems do offer flexible restructuring of a document collection; but the failure to recognize the central role of a controlled vocabulary in effective postcoordinate systems has led to the widely-held notion that keyword searching on full-text documents can replace traditional indexing structures. The effectiveness of full-text searching, however, is undermined by the very volume of documents available on the Web. The high percentage of irrelevant documents in retrieval sets generated by full-text searching dramatically underscores the failure of keyword (or pattern-matching) search engines to replicate or replace the collocating function of indexing structures such as thesauri, subject heading lists and classification schemes. Because natural language is riddled with homographs, search engines that employ pattern-matching techniques cannot ensure that the resulting retrieval set is conceptually relevant to the topic of the search query.

Collocation of conceptually-related documents depends upon the systematic application of a well-constructed representational language. Within the framework of an indexing language, whether a traditional classification scheme, a subject heading list or a thesaurus, the reference of a linguistic string is indexical and establishes a one-to-one relationship between a term and the concept to which it refers. Without recourse to an external index tool, however, the traditional classification scheme cannot address problems generated by synonyms and near synonyms. Indeed, the very complexity of Yahoo!'s classification scheme requires a free-text search to identify all potentially relevant classes. A thesaurus handles synonyms through the implementation of a syndetic reference structure that points the searcher to the authorized term for a particular entity, activity, property or attribute. But recent research carried out by Hert, Jacob and Dawson (In preparation; see also Jörgensen & Liddy, 1996) indicates that the use of a syndetic structure may not be appropriate in a hypertext environment where users expect to move directly from a term in the primary index either to the document itself or to a secondary index. They suggest that, to support ease of access through natural language terms and to minimize user frustration, developers of hypertext systems should consider implementing many-to-

ISSN: 2324-9773

Jacob, E., & Priss, U. (1999). Application of faceted classification structures. Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, 87-106. doi: 10.7152/acro.v10i1.12483

one indexing structures that provide multiple, redundant access points, each linked directly to the relevant document or secondary index.

Ideally, development of a many-to-one indexing structure would support natural-language access. Nonetheless, effective retrieval depends, ultimately, on a thorough understanding of how language is used within human institutions and activities -- within the knowledge domains and discourse communities that collectively constitute the sociocultural setting of a document collection (Hjørland & Albrechtsen, 1995). To support communication between the searcher and the indexer, an indexing language should both define domain concepts and indicate relationships that exist between those concepts – relationships that contribute both to the meaningfulness of individual concepts and to the conceptual organization of the domain.

Even if it were possible to establish an indexical, one-for-one correspondence between the documents to be indexed and the indexing language itself, problems would remain. Just as the meaning of a natural-language term must be understood within the context of the language game and form of life with which it is associated (Blair 1990), the full meaning of any one term can only be comprehended within the context of the conceptual relationships that inhere within the indexing language. Thus, when subject content is consistent across domains, the meaning of a term within one domain language may not be shared by related or competing domains. The lack of indexicality across domains is further compounded by the searcher's familiarity with the domain itself. Because the vocabulary of a given knowledge domain not only defines and orders the conceptual structure of that domain but serves also to distinguish levels of domain expertise (Jacob, 1994), the searcher's familiarity with a particular domain will impact ability to use the indexing language effectively.

Given these various linguistic and representational factors which directly affect retrieval, traditional implementation of an enumerative classification scheme will be less than effective in a dynamic and unstable environment such as the Web. Williamson (1998) points out that, while existing classification schemes have attempted to keep up with the constantly changing state of knowledge, they have not been successful in generating new structures that are able "to bridge gaps and to regroup concepts and ideas which are related to each other but spatially separate" (122) – structures that can overcome the problems inherent in current classificatory schemes. Williamson also observes that any changes in these schemes "have not been fundamental" (117) because developers have failed to address the continuing validity of the basic assumptions that underlie traditional classification schemes.

Rather than attempting to bridge the conceptual gaps in existing classification schemes, a more productive approach rejects the notion of a single fixed and universal pattern of reality in favor of the assumption that the social, cultural, historical and experiential origins of "reality" will necessarily generate a multiplicity of such patterns. Positing the existence of multiple patterns of reality allows for the potential instability of knowledge by viewing it not as an objective entity but as the emergent product of interpretation within one of many possible knowledge structures. The notion that knowledge is not absolute does not deny the possibility of knowledge *per se* but allows for a plurality of conceptual structures, each emanating within the framework of a particular knowledge domain. Acknowledging, too, that knowledge is not only malleable but also inherently fallible – that it is open to reanalysis, modification and correction within competing contexts -- will account for conceptual modifications which often occur where the boundaries of domain knowledge overlap (Jacob & Albrechtsen, 1999).

Replacing the notion of a fixed and universal reality with the assumption of multiple knowledge structures representing competing patterns of reality leads to rejection of the traditional understanding of representational structures as fixed and predetermined hierarchies of class relationships. Positing the existence of multiple patterns of reality supports the construction of polyhierarchical representational structures that can accommodate the potential overlap of phenomena across disciplines – or disciplines across phenomena -- that Williamson (1998) finds so problematic. It should be emphasized, however, that, while such pluralism will allow for the identification of divergent relational structures generated by varying sociocultural and historical experiences, it does not lead to an extreme relativism precisely because the occurrence of common phenomena, activities, attributes, theories and/or methodologies allows for convergence across domains.

One of the greatest obstacles to the development of an effective representational system – a system that will accommodate multiple patterns of reality within a polyhierarchical structure -- is the central role of the indexing language. In hypertext environments, the linking structure is inherently navigational. The failure of hyperlinks to function as an indexing system can be traced directly to the lack of consistent representation across documents and/or document collections (Girill 1991). While this may be attributed, in part, to the lack of a single controlled vocabulary that is shared across domains or across discourse communities, the problem is more fundamental. It is, in fact, a product of the failure to support indexicality at the conceptual level and can be traced to a reliance on *words* as descriptors -- the very problem that plagues many search engines. More importantly, perhaps, this is also a significant problem limiting the utility of traditional indexing languages and undercutting their ability to effectively represent a highly dynamic and increasingly interdisciplinary body of knowledge.

A controlled vocabulary or well-constructed indexing language rests on the basic assumption that, at some level, there is a single universal reality that can be mapped onto the indexing language such that each term will point to one and only one *concept* and each *concept* will be represented by one and only one term. Obviously, then, the failure of traditional indexing languages to adequately represent a dynamic body of knowledge is due, in part, to the assumption of universal reality – a single way of ordering the world. More importantly, however, such languages fail because their relational structure – the relationships established between classes -- is frequently driven by the linguistic basis of the notation. Rather than working from a set of concepts – the entities, activities, properties and attributes to be represented – the process of building a controlled vocabulary too often begins with the identification of the linguistic terms which will serve as the notation (Batty, 1989). Only after a set of terms has been collected will each term be associated with a given concept. The final step then involves building from the set of terms-cum-notation a single tree structure of hierarchical relationships. This process relies on the assumption that each word represents one concept but fails to account for the fact that the same word, when applied within differing contexts or domains, can refer to multiple concepts, the meanings of which are shaded by the sociocultural, historical or disciplinary context of each domain. Thus, just as the relational structure of DDC is inescapably hampered by the Procrustean nature of decimal notation and the restriction that all knowledge must be accommodated within a maximum of ten coordinate classes at each level within the hierarchy, an indexing language that begins with the accumulation of a set of words is inevitably constrained by the existing vocabulary.

The power and efficacy of an indexing language depend upon the indexical nature of the representations – the ability of the language to establish a one-for-one correspondence between a notation and the conceptual entity to which it refers. If the representational structure begins with identification of those linguistic strings which will constitute the notational labels, the system is inherently limited to defining one concept to be attached to each term. However, if such a structure begins with the identification of the set of concepts to be represented by an as yet unspecified notation, the indexing language will be better able to accommodate the variability of linguistic reference that frequently occurs across domains. By focusing on the concepts to be represented rather than the notational representations, it is possible to build an indexing structure capable of representation across domains – an indexing structure that is inherently flexible and responsive to the dynamic nature of human knowledge.

An approach that begins at the conceptual level and maps natural-language synonyms and near-synonyms to concept definitions will be able to accommodate variations in domain expertise – variations in language within the user population that reflect differing levels of familiarity with the knowledge domain itself. Thus, for example, *cello* and *'cello* are natural-language synonyms for the instrument correctly identified as *violoncello*. In a concept-based indexing system, each of these strings would be an authorized descriptor in the indexing language and would point to the same entity or concept: a user entering any one of these terms would retrieve the same set of documents. Unlike traditional indexing languages, the viability of this approach relies not on a controlled vocabulary but on the development of a structure of well-defined concepts, each of which is linked to one or more access terms. Development of a representational scheme based on a structure of concept definitions would support access across contrasting levels of user expertise and provide for greater

ISSN: 2324-9773

consistency of representation across indexers by linking natural-language synonyms to the appropriate concept within a controlled structure of concept definitions.

An alternative approach to organization

Review and analysis of problems incurred by traditional indexing methods point up the need to approach these problems from a different perspective. Rather than continuing to work from those conventional practices which have proven ineffective or problematic even within traditional environments, the emphasis must be on the identification of new approaches that will support the fundamental objectives of organization within the conceptual space of the Web. Emphasizing the primary objective of a controlled indexing language – the establishment of a one-for-one relationship between a descriptor and its referent rather than the explicit nature or source of such a vocabulary — and the problems incurred by existing languages shifts the focus away from the language itself and opens the way for identification of alternative structures that will effect the same, or similar, results. The need to facilitate retrieval across the boundaries of domain-specific languages points to the viability of a representational structure that moves behind language to the level of the concepts themselves. A viable alternative will therefore incorporate the basic principles that make controlled structures effective while accommodating the needs of a heterogeneous mix of users. Mapping natural language terms to corresponding concept definitions within a controlled structure of concepts and the relationships that inhere between them will provide flexibility within a representational system that can support access across natural languages and knowledge domains alike.

Williamson (1998) observes that, if a viable alternative to traditional structures were indeed possible, "such a system should be faceted in nature" (119). Although the theory underlying faceted structures is not new, faceted classification structures have not been widely implemented in traditional knowledge environments. When they have been implemented, such systems have generally been domain-specific. Faceting has been more widely applied in the development of thesauri, however, and the success of these ventures has tended to popularize the notion of facets. Indeed, recent popularization has made faceting a fad, but it has brought about a situation in which the function and structure of faceting have been misrepresented and/or misconstrued (e.g., Ellis & Vasconcelos, 1999).

Faceting is generally described as an analytico-synthetic process. Theoretically, a universe of knowledge, whether a domain of practice or a field of academic endeavor, is analyzed to determine the set of concepts that characterize the intellectual content of that domain. Once these central concepts – these entities, activities, properties and attributes -- have been specified, the paradigmatic variables, or elements, of each facet are identified (Priss & Jacob, 1998). Through a process of synthesis – through combination of the appropriate facet variables -- a representation is subsequently

ISSN: 2324-9773

generated for each document in the knowledge resource. For this reason, construction of a faceted structure is frequently approached as a bottom-up process that begins with collection of a set of terms specific to the knowledge domain. But while effective analysis must deal with those characteristics which typify the content of the target domain, it must do so within constraints imposed by the conceptual framework of the relevant domain. Effective analysis will therefore employ strategies that are both inductive (or bottom-up) and deductive (or top-down).

Because any concept can only be understood through its relationship to the full set of concepts which collectively constitute a given knowledge domain, the conceptual framework of a domain must consist of a set of baseline facets (Priss & Jacob, In press) in association with the set of relationships that hold between them. Thus the first step in developing an effective faceted language is necessarily top-down and involves analysis of the parameters of a knowledge domain and its particular perspective – the domain-specific world view or ruling paradigm -- that will influence the way in which that domain conceptualizes and organizes the phenomena of investigation. Once the domain perspective has been analyzed, analysis of domain phenomena will determine the principle characteristics that contribute to an initial set of baseline facets and the establishment of relationships between these facets.

While it is generally accepted that a single concept may be referred to by two or more terms, and that every term will necessarily refer to a concept, it is important to recognize that not every concept can be described by one term consisting of a single word or a phrase. If construction of the indexing language must rely on term acquisition as the principle means for concept identification, representation of domain concepts will be limited by, and therefore dependent upon, the existence of at least one term – one word or phrase -- that refers indexically to each domain concept. But because each concept can only be understood within the framework of its relationship to the other concepts which comprise the knowledge domain, concepts must be understood as independent of terms. In other words, the referent of a term – its *meaning* -- is neither inherent nor fixed but is arbitrary, depending as it does on immediate context and accepted usage within a particular knowledge domain. Terms must therefore be defined conceptually rather than linguistically: that is, each term must be treated as an element of linguistic notation that is mapped to, but nonetheless distinct from, the concept to which it refers (Priss & Jacob, In press).

This distinction between concept and term underscores the importance of the preliminary set of baseline facets and of the relationships established between facets, for it is within this initial framework that the full set of domain concepts will be constructed. More importantly, significant advantages accrue to an indexing language when the vocabulary of concepts is developed independently of the linguistic notation. As argued above, conceptual and/or linguistic overlap occurs across domains when the same concept is referred to by different linguistic strings in two or more domains dealing with the same phenomena, or, conversely, when a single linguistic string refers to two or more distinct concepts in separate domains.

Jacob, E., & Priss, U. (1999). Application of faceted classification structures. Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, 87-106. doi: 10.7152/acro.v10i1.12483

When two terms point to the same concept, whether they are synonyms within the same domain or occur in two separate domain vocabularies, both terms can be mapped to the same concept. For example, the technical term *violoncello* (used by the music scholar) and the common term *cello* (used by the general public) can be mapped to the same concept, thus accommodating the vocabularies of the domain expert and the non-expert within a single representational structure. Mapping two or more linguistic strings to a single concept can also provide access across natural languages within a single representational structure: for example, the German word *die Lüge* and the English terms *lie* and *falsehood* can be mapped to the same concept to support representation across natural languages. Conversely, when the same linguistic string refers to different concepts in two or more separate domains, each string can be disambiguated as part of the search process. For example, the system can prompt the searcher to distinguish between different referents – or meanings -- of the linguistic string *pig* (see Fig. 1):

Search term = *pig*
Select one:

- pig (domestic animal)
- pig (wild boar)
- pig (food : pork)
- pig (iron ore)
- pig (slang term : police)

Figure 1. Disambiguating *pig*.

When facets from different domains or languages are combined, there are two situations which illustrate the extreme points on a continuum representing combinatorial practices. In the first situation, many of the concepts would be identical; and those concepts which were not shared by both languages or domains would be easily accommodated within the joined structure. If a new object were discovered or invented, the object could be inserted at the logical point in the existing facet structure. For example, if a new planet were discovered, it would be inserted as one more element in the list of planets that comprise the baseline facet <planets>. If the facet <planets> were sorted by the characteristic "distance from sun", then the new planet could easily be inserted at the logical point in the list because its computed distance from the sun would necessarily be distinct from that of any other planet. The overall structure of the facet would not be altered because the characteristic “distance from sun" would still apply to all objects in the facet.

ISSN: 2324-9773

Jacob, E., & Priss, U. (1999). Application of faceted classification structures. Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, 87-106. doi: 10.7152/acro.v10i1.12483

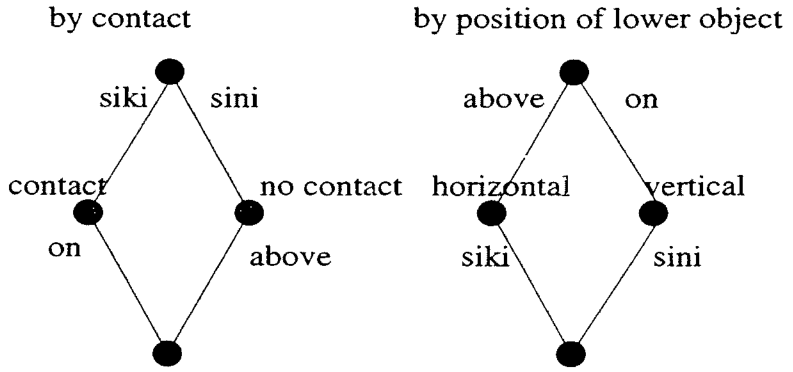

Fig. 2. Concept denotation based on distinguishing characteristic of facet.

In the second situation, two concepts from different domains or languages would overlap but the distinguishing characteristics for each concept would be quite different. For example, the English words *above* and *on* indicate the relationship between two objects: more specifically, each of these terms describes the situation where one object is positioned "higher than", or "over", the other object on a [hypothetical] vertical continuum. The characteristic that discriminates between these two terms is that of "proximity": *above* indicates no contact between the two objects, while *on* indicates that the two objects are in direct contact. In Mixtec (Regier, 1996), similar concepts are expressed by *sini* and *siki* except that "extension", rather than "proximity", is the relevant characteristic: *siki* is applied if the lower object is horizontally extended on the ground, while *sini* is used if the lower object is vertically extended. If the two facets were to be combined within a single conceptual structure, the English terms would be annotated at the top of the Mixtec facet, denoting the concept for which the distinguishing characteristic "extension" is not relevant (see Fig. 2). Similarly, the Mixtec terms would be annotated at the top of the English facet, denoting the concept for which the distinguishing characteristic "proximity" is not relevant.

According to Batty (1989), a bottom-up approach typifies formation of classes and development of a relational structure in a faceted structure. After defining the scope of the language and identifying sources for the vocabulary, terms that will comprise the language are collected on slips of paper (or the electronic equivalent) and grouped into classes. Only then are relationships between the classes established. In contrast, the double-edged approach advocated in Priss and Jacob (1998) – identification of a preliminary conceptual framework followed by an iterative process of mapping terms to elementai concepts within the existing framework -- provides for the construction of the indexing language within a conceptual framework derived from and applicable to the relevant domain(s). By combining bottom-up collection techniques with the

ISSN: 2324-9773

Jacob, E., & Priss, U. (1999). Application of faceted classification structures. Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, 87-106. doi: 10.7152/acro.v10i1.12483

top-down constraints imposed by an existing conceptual framework, faceted modeling of the target domain can accommodate potentially contradictory viewpoints within a single representational structure as long as intra-facet consistency is maintained at the conceptual level. This approach builds from identification of the various discourse communities that will have access to or participate in the knowledge resource and requires collection of multiple terminologies linked to the retrieval needs of a potentially heterogeneous mix of user groups. When the resulting faceted structure is stored within a relational database, changes produced by natural evolution of the constituent terminologies as well as modifications of the conceptual structure itself can be effected in an efficient and timely fashion by simply mapping the new terms within the existing conceptual structure. This ensures consistency of representation within the knowledge resource while allowing for currency of the access terminology and an inherent flexibility of the conceptual structure itself.

Application in a web-based environment

Implementation of a faceted language as a classificatory structure has traditionally used a fixed citation order to provide for systematic collocation of related materials. In the process of class formation through synthesis of facets and facet elements, a fixed citation order automatically generates a hierarchical tree structure of classes and supports the linear arrangement of records or physical objects. But the class structure imposed by a particular citation order can reflect only one perspective on the target domain -- a perspective that assumes universality in its imposition of a single conceptual organization. While a stable class structure is necessary for arrangement of physical materials in the collection, thereby supporting serendipitous exploration, the need for multiple avenues of access within electronic knowledge stores calls for the development of more flexible approaches -- approaches that will not only accommodate multiple perspectives of the resource materials but also provide access through the potentially diverse terminologies that frequently characterize a heterogeneous user population.

By storing faceted representations within a relational database, flexible generation of organizational structures can replace the rigidity inherent in a fixed citation order. Development of a set of combinatorial citation orders based on analysis of the profiles of constituent user groups or provisions for restructuring the knowledge resource according to a user-generated citation order will allow the system to accommodate a broader range of discourse communities and their various terminologies, levels of expertise and/or information needs. When implemented as a relational database within an electronic environment, a faceted indexing language can be used to generate a navigational structure of links for a hypertext Web site. A template for restructuring the Web site will be determined by the citation order; and the order in which facets are to be combined will determine the hierarchical and/or cross-referential links provided for the user.

ISSN: 2324-9773

One or more primary combinatorial orders may be established to represent the most common or expected ordering of facets and stored as one component of the faceted structure. These template structures will be used to order -- or re-order – the individual documents on the Web site as links across Web pages. Secondary or supplementary combinatorial orders will provide alternative organizational structures that can be implemented, as required, to provide alternative linking structures reflecting the needs and objectives of different segments within the user population. Thus, for example, the design of one combinatorial order might provide a "walking tour" that would guide a novice user through the most important or significant components of an electronic knowledge store. Whether a template is selected from a predetermined set or constructed by the individual user, provision of alternative approaches to organizing a collection of documents or document surrogates can be applied directly to the database to support dynamic [re-]structuring of the collection.

The successful application of this organizational approach is closely tied to the development of a concept-based representational system grounded within the framework of a well-defined conceptual structure. Words drawn from the terminologies of the contributing domain(s) will be mapped to the definitions that constitute this conceptual language, facilitating natural-language manipulation of the organizational structure by individual searchers. And, because the indexing language itself is faceted, it will be able to support, through the synthesis of existing facets and/or facet elements, representation of concepts that are not indicated by an existing linguistic term.

Practical application of this approach is currently underway with the development of a faceted indexing structure as the primary access tool for Indiana University's Knowledge Base [KB]. KB is a medium-sized collection of approximately 4500 computing-related documents that have been prepared by Indiana University's information technology staff [UITS] and are made available on the Web. These documents contain FAQ-like questions and answers, annotated lists of commands and manuals, as well as links to similar documents on the Web. KB was originally designed for the exclusive use of the academic community at Indiana University-Bloomington and its associated campuses across the state, but it has acquired a wide range of users and has received international acclaim. Documents are currently accessible through both a string-matching search interface and a basic hierarchical menu. A second-generation interface reflects a Yahoo!-like, enumerative approach to the organization of and access to the document store. This latter interface is nearing completion and will be available shortly.

Although the second-generation interface has not yet been released, development of a third-generation interface is already under way. This interface, which is intended to replace the Yahoo!-like hierarchical structure, implements the nontraditional approach described in this paper. The representational structure utilizes a faceted language of concept definitions to which natural-language terms are mapped. It also

ISSN: 2324-9773

Jacob, E., & Priss, U. (1999). Application of faceted classification structures. Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, 87-106. doi: 10.7152/acro.v10i1.12483

introduces the use of templates that will permit dynamic [re-]structuring of the document collection. The variations in citation order prescribed by these templates will be used to generate alternative linking structures for navigating through KB -- alternative systems of hyperlinks that will support the needs and objectives of different segments of KB's user population.

Conclusion

The nontraditional approach to organization described here demonstrates how conventional classificatory principles such as controlled vocabulary and fixed citation order can be effectively modified and adapted to the electronic environment. Analysis and evaluation of both the objectives and practical implementations of existing principles can lead to the development of alternative design frameworks -- frameworks that will generate new organizational structures that are at once stable yet flexible and able to accommodate the biases of various discourse communities. These alternative frameworks are made possible by addressing the fundamental discrepancies that exist between conventional organizational practices and the demands of a new environment such as the Web; by recognizing the need for more flexible and dynamic approaches to organization and rejecting the rigid class structure of traditional classificatory schemes; by refraining from the use of a linguistic basis for a representational scheme and electing, instead, to rely upon a more fundamental structure of concepts; and by foregoing accepted techniques for construction and implementation of a controlled vocabulary in favor of mapping natural language terms to an underlying definitional structure of controlled concepts.

The problems engendered by application of existing classificatory systems within the electronic environment are not insurmountable if information professionals will only step outside the traditional arena and look not to what is but to what could be. Eschewing the constraints imposed by traditional classification theory and adopting a more pragmatic approach to the problem of access will encourage an atmosphere hospitable to innovative approaches. Successful application of organizational principles to Web-based collections will require new perspectives on old problems -- perspectives that are unhampered by the technological capabilities of current search engines or the tenets of neo-Aristotelian philosophy and emanate, instead, from a thorough and unbiased analysis of both the ultimate objectives and current problems of Web-based organizational structures.

References

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders, fourth edition. Washington, DC: American Psychiatric Association.

Jacob, E., & Priss, U. (1999). Application of faceted classification structures. Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, 87-106. doi: 10.7152/acro.v10i1.12483

Barsalou, L. W. (1992). Cognitive psychology: an overview for cognitive scientists. Hillsdale, NJ: Lawrence Erlbaum.

Batty, D. (1989). Thesaurus construction and maintenance: a survival kit. *Database 12* (1), 13-20.

Blair, D.C. (1990). *Language and representation in information retrieval.* Amsterdam: Elsevier Science;

Blair, D.C. (1992). Information retrieval and the philosophy of language. *The Computer Journal,; 35*(3), 200-207.

Brier, S. (1996). Cybersemiotics: a new paradigm in analyzing the problems of knowledge organization and document retrieval in information science. In: P. Ingwersen and N.O. Pors (Eds.), *CoLIS 2: Second International Conference on Conceptions of Library and Information Science: Integration in Perspective; 1996 October 13-16; Copenhagen, Denmark, Royal School of Librarianship* (pp. 23-43). Copenhagen, Denmark: Royal School of Librarianship;

Bush, V. (1996/1945). As we may think. *Interactions, 3*(2), 35-46. Originally published in *Atlantic Monthly, 176* (1), 101-108.

DeBuse, R. (1988). So that's a book … advancing technology and the library. *Information Technology and Libraries, 7*(1), 7-18.

Ellis, D., & Vasconcelos, A. (1999). Ranganathan and the net: using facet analysis to search and organise the World Wide Web. *Aslib Proceedings, 51.* 3-10.

Foucault, M. (1970). *The order of things: an archaeology of the human sciences.* New York, NY: Vintage Books.

Frohmann, B. (1994). The social construction of knowledge organization: the case of Melvil Dewey. In H. Albrechtsen and S. Oernager (Eds.), *Knowledge organization and quality management: Advances in knowledge organization. vol. 4* (pp. 101-108). Frankfurt/Main: Indeks Verlag.

Girill, T. R. (1991). Extended subject access to hypertext online documentation. Part III: the document-boundaries problem. *Journal of the American Society for Information Science, 42*, 427-437.

Hert, C.A., Jacob, E.K., & Dawson. P. (In preparation). Evaluating indexing practice in the networked environment: an exploratory study.

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: domain analysis. *Journal of the American Society for Information Science, 46*(6), 400-425.

Jacob, E.K. (1994). Classification and crossdisciplinary communication: breaching the boundaries imposed by classificatory structure. In H. Albrechtsen and S. Oernager (Eds.), *Knowledge organization and quality management: Advances in knowledge organization, vol. 4* (pp. 101-108). Frankfurt/Main: Indeks Verlag.

Jacob, E.K., & Albrechtsen, H. (1997). Constructing reality: the role of dialogue in the development of classificatory structures. In I. C. McIlwaine (Ed.), *Knowledge organization for information retrieval: Proceedings of the 6th International Study Conference on Classification Research, 14-16 June 1997, London* (pp. 42-50). The Hague, Netherlands: International Federation for Documentation.

ISSN: 2324-9773

Jacob, E.K., & Albrechtsen, H. (1999). When essence becomes function: post-structuralist implications for an ecological theory of organisational classification systems. In T.D. Wilson & D.K. Allen (Eds.), *Exploring the contexts of information behaviour. Proceedings of the Second International Conference on Research in Information Needs, Seeking and Use in Different Contexts, 13-15 August 1998, Sheffield, UK* (pp. 519-534).

Jacob, E.K., & Shaw, D. (1998). Sociocognitive perspectives on representation. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology, vol. 33* (pp. 131-185). Medford, NJ: Information Today for the American Society for Information Science.

Jörgensen, C.L., & Liddy, E.D. (1996). Information zccess or information anxiety: an exploratory evaluation of book index features. *The Indexer, 20*(2), 64-68.

Liebscher, P. (1994). Hypertext and indexing. In R. Fidel *et al.* (Eds.), *Challenges in indexing electronic text and images* (pp. 103-109). Medford, NJ: Learned Information for American Society for Information Science.

McCarthy, L.P., and Gerring, J.P. (1994). Revising psychiatry's charter document: DSM-IV. *Written Communication, 11*, 147-192.

Nelson, T. H. (1994). Xanadu: document interconnection enabling re-use with automatic author credit and royalty accounting. *Information Services & Use, 14*, 255-265.

Olson, H. (1994). Universal models: a history of the organization of knowledge. In H. Albrechtsen and S. Oernager (Eds.), *Knowledge organization and quality management: Advances in knowledge organization, vol. 4* (pp. 101-108). Frankfurt/Main: Indeks Verlag.

Olson, H. (1996). Dewey thinks therefore he is: the epistemic stance of Dewey and UDC. In R. Green (Ed.), *Knowledge organization and change: Advances in knowledge organization, vol. 5* (pp. 302-312). Frankfurt/Main: Indeks Verlag.

Palmer, C.L. (Ed.). (1996). Navigating among the disciplines: the library and interdisciplinary inquiry. *Library Trends, 45*(2), 126-366.

Priss, U., & Jacob, E.K. (1998). A graphical interface for faceted thesaurus design. In E.K. Jacob (Ed.), *Proceedings of the 9th ASIS SIG/CR Classification Research Workshop* (pp. 107-118). Silver Spring, MD: American Society for Information Science.

Priss, U., & Jacob, E.K. (In press). Utilizing faceted structures for information systems design. In Knowledge: creation, organization and use: Proceedings of the ASIS 1999 Annual Conference, October 31-November 4, 1999, Washington, DC.. Medford, NJ: Information Today for the American Society for Information Science.

Regier, T. (1996). The human semantic potential: spatial language and constrained connectionism. Cambridge: MIT Press.

Williamson, N. (1998). An interdisciplinary world and discipline based classification. In W. M. el Hadi, J. Maniez, & S. A. Pollitt (eds.), *Structures and relations in knowledge organization: Proceedings of the Fifth International ISKO Conference, 25-29 August 1998, Lille, France* (p. 116-124). Würzburg, Germany: Ergon Verlag.

ISSN: 2324-9773