KO Reports

12th American Society for Information Science & Technology, Special Interest Group/Classification Research. Classification Research Workshop

Alexander Justice

University of California, Los Angeles

The 12th classification research workshop of the American Society for Information Science & Technology, Special Interest Group/Classification Research 2001 was held as part of the 64th ASIS&T Annual Meeting, November 2-8, 2001 in Washington, D.C. A proceedings preprint was distributed to registered participants. The workshop papers will be published in final versions in mid-2002 by Information Today (http://www.infotoday.com) as *Advances in Classification Research. Volume 12.* In addition, transcripts of the discussion and commentary associated with the paper presentations will be added to that volume. The 12th SIG/CR workshop opened with a keynote address delivered by Dagobert SOERGEL, College of Information Studies, University of Maryland. The address was entitled *The many uses of classification: Enriched thesauri as knowledge sources*.

The presentations at this year's SIG/CR Workshop progressed in scope from general to specific, as the Workshop Program, indicates:

Workshop Program

Keynote Address

Dagobert SOERGEL, College of Information Studies, University of Maryland
The many uses of classification: Enriched Thesauri as knowledge sources

Session 1

Joseph T. TENNIS, Information School, University of Washington
Layers of meaning: Disentangling subject access interoperability

Uta PRISS, School of Library and Information Science, Indiana University
Multilevel approaches to concepts and formal ontologies

Session 2

Allyson CARLYLE and Sara RANGER, Information School, University of Washington
Facilitating retrieval of fiction works in online catalogs

Stephanie W. HAAS[1], Debbie A. TRAVERS[1], Anna WALLER[2], Brian HILLIGOSS[1], Molly CAHILL[1], and Patricia PEARCE[1,3], University of North Carolina at Chapel Hill, [1]School of Information and Library Science; [2]Department of Emergency Medicine, School of Medicine; [3]School of Nursing
Defining clinical similarity among ICD-9-CM diagnosis codes: diagnosis cluster schemes

Session 3

Brigitte ENDRES-NIGGEMEYER[1], Bernd HERTENSTEIN[2], Claudia VILLIGER[1], and Carsten ZIEGERT[2], [1]Fachhochschule Hannover/University of Applied Sciences, Department of Information and Communication; [2]Medizinische Hochschule Hannover/Hanover Medical School, Department of Hematology and Oncology
Constructing an ontology for WWW summarization in bone marrow transplantation (BMT)

The overarching theme elaborated in the presentations was the improvement of information retrieval by means of more effective approaches to classificatory activities on a number of fronts: controlled vocabularies, bibliographic systems, and ontologies, This theme might also be described as working towards a more meaningful information retrieval via more sophisticated approaches in classificatory theory and technique.

In *Layers of meaning: Disentangling subject access interoperability*, Joe TENNIS challenged the conceptual boundaries surrounding the issue of useful *shifting* between controlled vocabularies, classifications and thesauri. TENNIS evaluated and found inadequate the existing forms of mapping and switching systems developed so far. He posited an ideal situation of *interoperability* in which users would have their ability to recall information across collections powerfully enhanced. The obstacle to this ideal is of course the great variety of both universal and specialized subject access schemes upon which users of these diverse collections depend. TENNIS proposed that "a mechanism must be built that allows the different controlled vocabularies to communicate meaning, relationships, and levels of extension and intension..." The paper first set out in brief the path of development toward this longstanding goal of compatibility, reviewing mapping between classifications and then switching languages such as the Information Coding Classification and the Broad System of Ordering. TENNIS introduced the idea of semantic layers, or layers of meaning, as a key to understanding a) where past solutions fell short and b) where new solutions must find adequacy. Layers, in Tennis's view, are used by interoperability mechanisms to control his identified variables of "meaning, relationships, and levels of extension and intension." The use of only one layer to achieve interoperability inevitably failed to address all the problems facing the task, problems of subject overlap, specificity, degrees of pre-coordination, and relationship structures such as hierarchy and synonymy. He identified necessary semantic layers as those of concepts, subjects, and classes. Disentangling these layers before proceeding with the design of subject access interoperability systems is the goal Tennis set out, warning that the set of semantic layers he enumerated was only the first stage of layers in a "multi-layered conceptual framework" that should be developed in pursuit of the overall goal of subject access interoperability.

In *Multilevel approaches to concepts and formal ontologies*, attendees were presented with the argument that formal ontologies could benefit from a new view of the relationship between formal, or symbolic, representation and "fuzzy or category-based approaches to representation." Uta PRISS explored the division, found in various disciplines, between "formal concepts" and "associative concepts" and their implementation in formal and associative representations of knowledge or cognition, especially in the fields of knowledge organization and artificial intelligence research, respectively. She argued that both approaches presented limitations which might be overcome if they were integrated, and cited research which indicates that in the human mind, both approaches appear to be utilized if not integrated. If this is the case, it would make it incumbent on classificationists to work toward a similar and helpful integration in their systems.

PRISS initially reviewed and contrasted emergent structures, arising out of collective activity and not subject to direct control, and designed structures in which direct control over the system is paramount. To exemplify these differences, familiar WWW IR services were ranged in a spectrum from Google to Yahoo! Thus, Google represented the use of emergent structures centered on the dynamics of "the WWW linkage structure itself," Lycos and Altavista represented the use of emergent structures centered on the dynamics of "natural language processing techniques" and Yahoo! represented a totally designed structure with "no room for emergent structures at all." Emergent structures thus exemplify associative approaches in contrast to formal approaches or the designed structure. PRISS then pointed to possible combinations of associative and formal approaches, applied in that order, such as partially exemplified by the WWW IR service Northernlight, which offers "an automatically generated folder hierarchy," or an associative approach augmented by a formal approach. Further exploration of associative and formal approaches used examples of cognitive contexts to highlight the use of both in "human rationality." The activity of "concept formation and definition" may move "seamlessly and unconsciously" in such contexts. Thus humans think of birds as flying animals even though not all birds fly, formalizing their definitions only when necessary. "The two levels of formal and associative approaches are often complementary. Human cognitive acts are usually neither solely associative nor formal but instead a combination of both." In addition, the possibility of symbolic representation of concepts creates further levels that affect the associative and the formal. Symbols, systems of symbols (e.g. natural language) and the external world itself all form levels of representation that must inform an exploration of the interaction between the associative and the formal on behalf of knowledge representations or ontologies. PRISS therefore advocated a "multilevel approach" for ontologies. This could "represent non-symbolic knowledge in 2- or 3-dimensional schematic simulations, which would be mapped to associative concepts using gestalt principles of perception."

PRISS then reviewed "dynamic interactions between associative and formal levels" to press the case that ontologies would benefit from the combination of both associative and formal approaches to their task. Language itself, depending on one's point of view, functions as either an associative or a formal system. Formal logic may produce "emergent structures...that cannot be explained within the original system." Computer games take players from open to complete structures, whence the player begins again on a new level with a new but more complicated open structure. Visual representations such as diagrams and maps are used alongside formal logic or verbal instructions in both education and daily life, and many "mental tasks...are best tackled" by a strategy of alternating or shifting between associative and formal approaches, most notably writing. PRISS concluded with the hope that formal ontologies, by "incorporating associative structures," would in return receive an exponential increase in their capacity to effectively represent knowledge.

In *Facilitating retrieval of fiction works in online catalogs* the spotlight shifted to the arena of bibliographic control and the absence so far of "systematic retrieval and meaningful display of bibliographic records" of those works of fiction whose popularity and canonicity have vastly multiplied their editions and adaptations. Allyson CARLYLE and Sara RANGER demonstrated how the situation could be rectified by almost entirely automatic means. A well-designed technique of classification utilizing already existing data in the machine readable records should permit the automatic creation of work classes or sets, and retrieval and display of these sets, rather than individual records in hit-or-miss fashion, was proposed as the norm.

The rationale for the project was reviewed in light of existing studies on catalog use. In the case of "well-known and frequently sought works" published in many editions, several problems present serious obstacles to users whether the search were conducted by author, author and title, or title. These manifest to the user as the apparently disorganized presentation of retrieved records (caused by alphabetical order title display), as incomplete retrieval sets (in the case of fiction especially, where many editions often have many varying titles), and as dauntingly large retrieval sets. CARLYLE and RANGER also reported that related research shows promise for the meaningful and automatic arrangement of result sets through use of some attributes already present in MARC records.

In order to more thoroughly identify the attributes likely to be utilizable in such automated bibliographic organization, CARLYLE and RANGER selected four representative works of fiction by the likes of Dickens, Stevenson, Dumas and Alcott. Bibliographic records representing editions (and editions only) of these works were then subjected to scrutiny, manually, in order to clearly identify what information in which fields could be put to use. The records examined were obtained from two sources, the OCLC Office of Research and the OLUC. The former provided English-language edition records garnered previously through automatic methods; the latter source was painstakingly searched by author and title, and cross-checked against the NUC and authority records. In addition, CARLYLE and RANGER set out a definition of classification "as a process comprised of two interrelated actions," i.e. the use of specified attributes in order to identify the "work class" to which a record should belong, followed by the "assembly" of all such records.

Thus the results of CARLYLE and RANGER'S investigation were set out in three stages: attribute identification, automatic attribute identification, and automatic clustering results. The authors identified author name, standard title and LC classification number as utilizable in discovering and identifying works. Automatic identification and classification of records as a work set could then employ "single attributes and attribute combinations" such as name plus title where name and title appeared in separate fields (MARC 100 and 240, 245, etc.), or name plus title where name and title appeared in a single field (MARC 700), or LC class number alone (MARC 050 or 090). CARLYLE and RANGER described a theoretical program of automatic discovery of attributes by systematic harvesting of "sets of records" from the Library of Congress Name Authority File (NAF), which would then be automatically cross-checked against further attributes within each record. Automatic clustering was simulated by the researchers through a process of manual analysis. They concluded that adequate automatic clustering must identify and exclude, that is, identify the records that belong to the work set, and yet exclude those that do not. Related works were found to be excludable by the presence of a variety of attributes in their MARC records, which CARLYLE and RANGER believed could form the basis of a more refined set of criteria for the exclusion function of the automatic clustering process.

CARLYLE and RANGER indicated a success rate of 86 to 98 percent for "a small number of automatically identifiable attributes used together in attribute combinations to automatically classify work records," a rate which would be higher were it not for the non-English origin of one work among the four selected. This strong success rate seemed clearly to call for more investigation in this area so as to provide the basis for developing specific improvements to online catalogs. The goal of such improvements would be a far more useful "retrieval and display" of many-edition works of fiction.

In *Defining clinical similarity among ICD-9-CM diagnosis codes: Diagnosis cluster schemes,* Stephanie W. HAAS *et. al.* address a difficulty encountered with an application of the International Classification for Diseases (9th edition, Clinical Modification or ICD-9-CM). While this classification was designed to deal with mortality and morbidity, it is today used for "reimbursement and reporting diagnosis" among other information gathering tasks. Users have found its highly granular quality an obstruction for these new purposes. In attempting to understand why patients visit an emergency department, for instance, there appear to be too many classification terms (1,600 codes assigned in the cases of some 5,000 patients), making the raw data collected in emergency departments – the "what" of a visit – unsuitable in answering the "why." As a possible solution, this interdisciplinary team at the University of North Carolina at Chapel Hill examined diagnosis clustering schemes in medical information and questions about them.

These questions boil down to the place of diagnosis clusters among "myriad representations" of medical information, the principles to be used to define such clusters, and the portability of cluster schemes from one use to another. The context for approaching these questions is emergency medicine (EM), and the focus is on the process of grouping or clustering existing codes (such as those of the ICD-9-CM) "for a specific purpose." This process is of course a form of classification, and depends on the selection of those characteristics that will guide the grouping of terms. Therefore this inquiry appears to involve the classification of a classification, prompted by the need to represent new information using an existing standard. The characteristics of interest here are those that exhibit "clinical similarity." This is problematic, however, because different healthcare practitioners may not view clinical similarity in the same ways. Indeed, "many have commented that [the ICD-9-CM] hierarchy of codes does not support aggregation for their purposes."

The key question in this case is whether or not diagnostic cluster schemes can be found or developed for EM. Three such schemes derived from ICD-9-CM were therefore analyzed, those relating to outpatient family medicine, ambulatory internal medicine, and hospital inpatient stays. The researchers did not find these schemes to rely unambiguously on the same definition of "clinical similarity." Nevertheless they were applied to "a sample of ED final diagnoses," and this revealed that "difference in purpose and specialty affected their design." None of the three was found to be adequate for EM unless modified, and in fact a scheme for EM is in development. The issue of "clinical similarity" remains very important as a research question in the production of a diagnosis cluster scheme for EM, but is held as one of two main foci, the other being "practical issues" about the size and scope of the clusters themselves.

In *Constructing an ontology for WWW summarization in Bone Marrow Transplantation (BMT)* Brigitte ENDRES-NIGGEMEYER *et al* presented the progress to date in the production of a domain specific, grounded ontology for medical knowledge representation in Bone Marrow Transplantation (BMT). This ontology will play a key role for physicians, "summarizing agents and ... other system participants" and is being designed especially to enable automatic summarization from documents existing on the World Wide Web. It is intended to be a "dense representation of domain knowledge" which for any given concept "encompasses statements of relevant knowledge about it." Hence its web site title is "Summarize It in Bone Marrow Transplantation" or "SummIt."

The main functions of the ontology are query scenario formulation, text passage retrieval, and summarizing. The ontology, based on an XML server, is to play its roles during at least these three phases of the summarization system. In other words, it will be made use of by the agent initiating the query and search to begin with, as well as during the following processes of evaluating and then summarizing retrieved WWW documents. The ontology and its methodology appears to be unique to its environment, and not all of its details translate well to the brief report; readers may wish to inspect the web site referred to by the authors: http://summit-bmt.fh-hannover.de

Key features of the development of this BMT ontology are groundedness, user-centeredness, a strict adherence to formal logic and an avoidance of fuzzy approaches, and an inductive system design approach in which form follows function. In this case groundedness is taken to demand use of domain texts for accurate knowledge modeling, especially those texts exemplified by the domain's core journals. The designers also made clear their aim to ensure that the users are also "responsible co-authors" as the system under construction will depend upon the ontology, and would be severely undermined by insufficiently accurate terminology. The designers further explained the need for precision in the life-and-death world of BMT. "In medicine many statements are valid only if the limits of their scope are respected. Perhaps more than elsewhere, we have to represent preconditions..."

According to the designers, the BMT ontology is at present about one-fifth complete. They asserted the likelihood that it would change very little in its design from this point forward as part rationale for presenting their research at this juncture. The design of the BMT ontology is based on thesaurus construction principles as understood in LIS, as well as grounded theory as understood in sociology. The ontology was therefore described as being built up inductively, where "concepts are justified by and connected with their evidence, found almost always in text." Expert recommendation led to the identification of two primary bodies of texts, BMT papers published in *Blood* "a core journal of the domain" and educational papers derived from the Association of Hematology. Actual ontologies were produced based on each paper in these source groups, that is, the complete listing of "concept occurrences" in each paper in question. Later in the process, "more concise statements" were recorded, as well as "predicate logic expressions." Thus the BMT ontology is built up from textual analysis of the domain. From the other end, the designers had established deductively the "upper model" of the ontology (Process, Thing, Element categories) by modifying MeSH where possible (the resulting taxonomy was largely their own, however) and by comparing with other thesauri systems in medical science. This "upper model" provides the general categories to which the inductively derived specific concepts naturally belong. Categories of Processes, for example, are therapy, laboratory test or technique, imaging process.

In addition to deriving and organizing concepts as any thesaurus would do, the BMT ontology, as described by ENDRES-NIGGEMEYER *et al*, also develops statements about the medical knowledge behind each concept. In doing so it moves from "textual context to first-order context expressions." These context expressions are not stored in the individual concept records but in an additional central database. Context expressions are generated from two kinds of propositions: context propositions and core propositions. Thus the context proposition *priortherapy (bone marrow transplantation, , relapse)* and the core proposition *treatmentOption (, second bone marrow transplantation)* unite to form the context expression *ist* (*priortherapy (bone marrow transplantation, , relapse), treatmentOption (, second bone marrow transplantation))*. Stated abstractly, "context expressions assert that the proposition p is true in the context c: *ist (c, p)*."