ASIS&T SIG/CR 2000. Classification for User Support and Learning

The 11th ASIS&T SIG Classification Research Workshop (ASIS&T SIG/CR 2000, Dagobert SOERGEL, Padmini SRINIVASAN and Barbara KWASNIK, co-chairs) was held on Sunday, Nov. 12, 2000 as part of the 62nd ASIS&T Annual Meeting. As the result of a highly competitive selection process, it brought together papers under the theme **Classification for User Support and Learning**. The program is given in Figure 1. Some of the papers are available on the Workshop Web site at http://uma.info-science.uiowa.edu/sigcr/. Final versions of the papers will be published mid-2001 by Information Today as *Advances in Classification Research. Volume 11.*

Main program

Introduction and foundation

David JONASSEN (invited), School of Info Science and Learning Technologies, Univ. of Missouri
Knowledge is complex: accommodating human ways of knowing

Session 1. Developing user-oriented classifications

Marianne Lykke NIELSEN, The Royal School of Library and Information Science Institute of Information Studies, Denmark
Domain analysis, an important part of thesaurus construction. Methodologies and approaches

Stephanie W. HAAS and Carol A. HERT, School of Info. and Libr. Sci., Univ. of North Carolina, Chapel Hill, and School of Info Studies, Syracuse Univ.
Terminology development and organization in multi-community environments: the case of statistical information

Session 2. Classification in the user interface

Nina WACHOLDER[1,2], *Judith VENUTI*[1,3], *Michael KRAUTHAMMER*[4], *and Pat MOLHOLT*[1] Columbia University, [1]Off. of Scholarly Resources; [2]Ctr for Research on Information Access; [3]Dpt of Anatomy and Cell Biology; [4]Dpt of Medical Informatics
Accessing and browsing 3D anatomical images with a navigational ontology

Susan DUMAIS (invited) and *Ed CUTRELL*, Microsoft Research, and *Hao CHEN*, Univ. of Calfornia, Berkeley
Use of classified displays of Web search results

Idea mart

Marcia Lei ZENG and *Pat MOLHOLT*, Kent State U. and Columbia U.
Knowledge organization scheme for cross-cultural and cross-language information systems -- issues and challenges

Yi-Fang WU, School of Info. Science & Policy, State University of New York at Albany
Automatic concept hierarchies development: A revised subsumption approach

Tony TSE, College of Information Studies, U. of Maryland
Identifying and characterizing a Health Consumer Vocabulary

Laura SLAUGHTER, College of Information Studies, U. of Maryland
Interfaces for understanding: Improving access to consumer health information

Elin JACOB, Elizabeth DAVENPORT, Uta PRISS
The world of Pokémon: A dynamic ecological classification system

Elisabeth DAVENPORT, Napier U. Business School, *and Howard ROSENBAUM*, and *Uta PRISS*, SLIS, Indiana U.
Ethological classification: a model for ordering the commercial workplace that draws on collective practice

Peiling WANG, University of Tennessee
Comparing cognitive maps using graph algorithms

Stephen PALING, School of Info. Studies, Syracuse U.
Information cartography: A proposed model for access to heterogeneous end-user databases

Winfried SCHMITZ-ESSER, U. of Appl. Sci., Hamburg, Germany
SERUBA - A new search and learning technology for the Internet and intranets

Session 3. Automatic creation of representations

Susanne M. HUMPHREY, Thomas C. RINDFLESCH, and Alan R. ARONSON, Lister Hill National Center for Biomedical Communications., National Library of Medicine
Automatic indexing by discipline and high-level categories: Methodology and potential applications

Hidetsugu NANBA, Noriko KANDO, and Manabu OKUMURA, Japan Adv. Inst. of Sci.& Technol. and Nat. Inst. of Informatics
Classification of research papers using citation links and citation types: Toward automatic review article generation

Alejandro JAIMES[*], *Ana B. BENITEZ*[*], *Corinne JOERGENSEN*[¥], *and Shih-Fu CHANG*[*], [*] Columbia University, [¥]State University of New York at Buffalo
Experiments in indexing multimedia data at multiple levels

Jack ANDERSEN, Royal School of Library and Info Science, Denmark
Document Theory and Knowledge Organization. An Approach based on Epistemology and Sociology of Knowledge.

Figure 1: Program

The lead-off speaker was David JONASSEN, Distinguished Professor, School of Information Science and Learning Technologies, University of Missouri. He provided a perspective underlying the workshop in his talk *Knowledge is complex: accommodating human ways of knowing*. The paper's main message was that we need classifications for the different kinds of knowledge that users hold and seek, particularly types of knowledge that are intimately tied to doing. The types of knowledge he covered are:

1 Ontological (Domain-specific) knowledge types
  1.1 Declarative knowledge
  1.2 Structural knowledge
  1.3 Conceptual knowledge
2 Epistemological (task-specific) knowledge types
  2.1 Situational knowledge
  2.2 Procedural knowledge
  2.3 Strategic knowledge
3 Phenomenological knowledge types
  3.1 Tacit (implicit) knowledge
  3.2 Compiled (automated) knowledge
  3.3 Sociocultural knowledge
  3.4 Experiential (episodic) knowledge
  3.5 World knowledge

Session 1 showed a wide range of methodological tools for constructing thesauri / classifications / ontologies.

In *Domain analysis, an important part of thesaurus construction: Methodologies and approaches*, Marianne Lykke NIELSEN introduced and illustrated domain analysis, a multi-pronged method to discover users' task approaches, resulting information needs, conceptual frameworks, and terminology as the basis for constructing a truly user-oriented thesaurus, exemplified by a thesaurus for a pharmaceutical company. Domain analysis focuses on the following factors:

- the nature of the professionals (background, work tasks, information needs, information use, language use, searching behavior, search problems),
- the subject field (topics, concepts, vocabulary),
- the literature (type, level, quantity), and
- the available resources for indexing and thesaurus construction (competence, time).

NIELSEN used the following methods:

- Group interviews to obtain an understanding of the work domain and its users.

- Content analysis and discourse analysis of user requests to investigate from which perspective and aspects the users approach the particular subject field.
- word association test to identify language use and approaches to the subject field.

In *Terminology development and organization in multi-community environments: the case of statistical information*, Stephanie W. HAAS and Carol A. HERT presented a conceptual framework and methodology for discovering concepts, concept relationships, and terminologies used by different user communities concerned with the same subject matter, in the example statistical data. Their methodology consisted of three parts:

- Constructing a conceptual map of the expert terms.
- Expanding the expert terms by adding synonyms from a thesaurus.
- Identifying user terms from query logs and matching them with the expanded expert terms, automatically where possible and manually where necessary.

The results can be used to construct a cross-walk from the search terminology of "lay" users of a statistics Web site to the terminology used by the experts who create the Web site.

Session 2 dealt with using classification for enhanced searching and display. It brought to the fore issues of user interaction with and navigation within graphical and text-based information and the role that a thesaurus or structured display can play in these.

In *Accessing and browsing 3D anatomical images with a navigational ontology*, Nina WACHOLDER *et al.* presented the Vesalius Anatomy Browser (www.cpmc.columbia.edu/projects/vesalius), an elegant system for searching and displaying anatomical images that is based on an ontology of body systems and body parts using several types of relationships as shown in the following illustration (Figure 2):

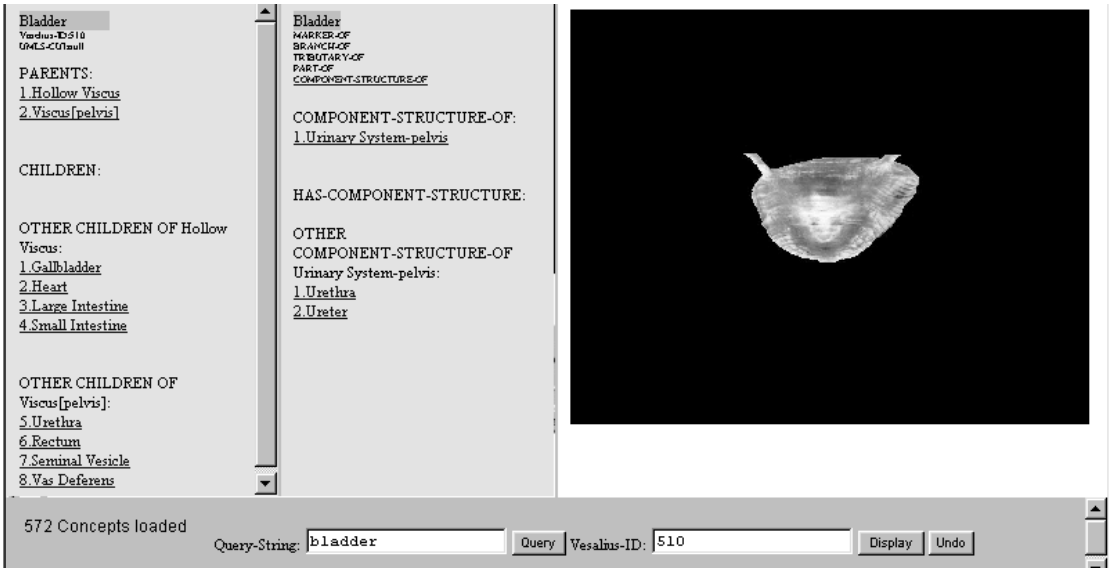

Figure 2: Screenshot from the Vesalius Anatomy Browser

The presentation elucidated the problem of learning how to deal with massive amounts of data in a visual display in a manner that is useful and comprehensible. The solution taken by the project team was to add explicit conceptual information to the system; otherwise the displayed information was only meaningful to an expert, in this case an anatomist. Their "navigational ontology" supports restricted inference and restricted relationships. In this case, the anatomically significant relationships are conceptual, functional, and spatial. The two major types of relationships in the system are *part-of* (component-structure) and *is-a* (taxonomic). Is-a and part-of, however, are not simple relationships; in a visual environment their complexity becomes more obvious, because there are matters of granularity and scale and multiplicity of types. For example, in addition to component struc-

ture there are other kinds of part-whole relationships, such as region and marker relationships; indeed, some things only makes sense as part of a larger structure, and one problem is that those things simply can't be made smaller.

One important point that holds implications for thesaurus display is that not all combinations of structures have *names*, therefore the system was designed to show relationships among structures, enabling the user to choose *non-named* sets or groups.

One question is whether, in a visual navigational ontology, there are other relationships besides is-a and part-of structure that need to be added. Wacholder stated that currently, more spatial relationships are needed, such as "nearby" or "part of two systems," enabling the user to search on more combinations. Adding 3-D creates another set of relationships and the major issue becomes one of classification, and not interface design.

What other non-visual relationships may need to be added in the domain of anatomy? And when does one start adding other relationships outside the scope of anatomy but of interest in a wider medical research domain, such as similar biochemical processes? One can foresee, far in the future, a system encompassing many types of knowledge about the functioning of the human body and capable of displaying not only non-named "things" but also facilitating new discoveries by relating previously unrelated structures, processes, and outcomes.

In *Use of classified displays of Web search results*, Susan DUMAIS *et al.* presented empirical evidence that classified displays of Web search results are indeed useful; they perform better than simple ranked-list displays both for user tasks and in user preference. Here is an example of a category display (abridged and simplified, all titles are hyperlinks):

Query: jaguar

Automotive
Jaguar Club of Florida
A&L Luxury Car Center - Jaguar Main Page

Computers & Internet
Atari Jaguar System
Jag-Lovers Jaguar Cars Windows Wallpaper page

Entertainment & Media
The Jaguar Phot Gallery

Travel & Vacation
Welsh Jaguar Classic Car Museum

Automatic techniques are used to map search results to a pre-established scheme of categories; the advantages of this method are that a user can quickly know the structure of the information and that the display is easily understood. In contrast, clustering techniques are used primarily to discover structure; in a retrieval interface this technique is slow and it is hard for the user to understand what the groups are.

The user study confirmed the advantages of a category display over a list display, both in terms of search times and user satisfaction. Interestingly, the researchers found that users could tolerate some ambiguity and "fuzziness" in the display. Items could be in multiple places (they could be placed in up to 13 categories); subjects noticed this feature and liked it. The automatic classification is not perfect; users noticed errors but were not bothered by them.

User comments regarding errors brings up the question of how "perfect" a classification process should be. While a large amount of error will cause users to distrust a system, greater accuracy requires larger amounts of time (and thus money). To what extent should we strive for "perfection" in classification of *heterogeneous* documents in very large databases (e.g., the Web)? As the standard for improvement is generally taken to be an increase in precision of 10%, techniques that create only small incremental improvements may not be worth the time and money invested in their creation. Related to this concern is the idea that in creating a classification system tailored to the needs of a particular user community, we reinforce domain boundaries, whereas classification of large heterogeneous collections would seem to need some permeability across these boundaries. As Dumais, Cutrell, and Chen's work shows, improvement in display of search results can minimize the impacts of "imperfect" categorization.

In *SERUBA - A new search and learning technology for the Internet and intranets*, Winfried SCHMITZ-ESSER gave a preview of a Web search system that uses a thesaurus with a rich set of relationship types to help the user explore her search topic. The relationships used are

Abstract/generic	**Cause/effect**
Partitive (physical and theoretical)	**Beneficial**
Partitive (habits, law and jurisdiction)	**Detrimental**
Partitive (geographical, topographical)	**Process applied**
Instrumental	**Derivative**

When the user enters a search term, the system uses synonym relationships to identify the corresponding concept and then displays other concepts in an array arranged by type of relationship, as in this abridged display (each referenced concept is a hyperlink):

Telecommuting	
is narrower concept of	labor
	new ways of working and living
is broader concept of	mobile telecommuting
	alternating telecommuting
is instrumental for	organizing work effectively
causes	flexible work time
	energy conservation
is beneficial for	virtual organizations
	combining family and work
is detrimental to	face-to-face contacts
by instruments	telecommuting workplaces
	online technology

The system displays results using its Basic Semantic Reference Structure, a frame whose slots can be seen from Figure 3.

Rules for assigning journal descriptors were developed based on statistical association of document features, such as title words, with journal descriptors assigned to documents in a training set. The rules for assigning semantic types rely on a more complex indirect method.

In *Classification of research papers using citation links and citation types: Toward automatic review article generation*, Hidetsugu NANBA *et al.* presented a tool box for the automated or computer-assisted generation of reviews based on analyzing citation relationships. The three tools would each be useful individually: a tool to identify and demarcate areas in a document that are concerned with reference to and discussion of a cited document; a tool for determining the type of citation relationships; and a tool for automatic classification of a document or document passages based on typed citation relationships. The citation area tool starts from the sentence containing the citation and adds sentences preceding or following based on the occurrence of cue words that indicate text cohesion. The citation type tool is also based on

What?	**Who?**		**Event?**	**Where?**	**When?**	**How?**
Universal, concept, theme	**Person**	**Corporate body**	**Name of event**	**Space**	**Time**	**Aspect**
General manager	Mike Osborne	Asia Trading Co., Vancouver		Canada	> 1998-11-1	Definition
Planting of St. John's trees		Ministry of Agriculture, Lima	El Algarrobo project	Peru	> 1984	Propagation

Figure 3

Session 3 dealt with automated methods to create the knowledge structures necessary for good user support.

In *Automatic indexing by discipline and high-level categories: Methodology and potential applications*, Susanne HUMPHREY *et al.* developed a system for automatically indexing documents with broad descriptors that express the general nature and orientation of the document and thus are useful complements to specific descriptors. Two types of broad descriptors were assigned: a broad scheme used at NLM to categorize journals by subject (127 Journal Descriptors, such as *Drug therapy*, *Antibiotics*, or *Pulmonary disease (specialty)*) and the 134 semantic types defined in the Unified Medical Language System, such as *Spatial concept*, *Therapeutic or preventive procedure*, or *Medical device*.

cue words to assign a citation to one of three types: show other researchers theories and methods; point out problems or gaps in related works, and other. The paper discusses both word-based and citation-based approaches to automatic classification.

In the middle of the day, an "idea mart" was held. It was devoted to extensive discussion of emergent research ideas or projects in small groups in five parallel sessions covering two topics each. This experiment turned out very well, producing many useful suggestions for the research of the presenters.

The second part of this report discusses themes that emerged from the papers and discussions. Some themes are clearly tied to one paper while others emerged in several papers. An overview follows.

Some themes in classification research

Theme 1.	Expanded use of classifications
Theme 2.	Requirements for diversity in classification
Theme 3.	The quest for unity. Multi-purpose classifications, reuse
Theme 4.	Types of knowledge covered in classifications
Theme 5.	Orientation of classification: Users' conceptual structures or intrinsic logic of the domain
Theme 6.	Types of relationships in a thesaurus / classification / ontology
Theme 7.	Display and user interaction issues
Theme 8.	Practical issues

Theme 1. Expanded use of classifications

Several presentations call into question the restricted uses that classification schemes have played, being used primarily for the organization of information for retrieval. Other roles that need to be explored more fully include roles in learning (e.g. the use of the visual anatomist for training and education), exploration and browsing, creativity, discourse, problem solving, and information

One question that emerged from these discussions is: How do we build classification systems that would enable us to discover and see relationships that have not yet been established?

Theme 2. Requirements for diversity in classification

Classifications should serve a given purpose for a given user community. Language – terms and their relationships with each other – is complex; it shows differences not only across domains but also across user groups in the same domain.

There are many sources of diversity in the design of classifications.

Sources of diversity

- Knowledge is complex (title of the first talk but an underlying theme of all).
- Many types of knowledge
- Many (discourse) communities / communities of use
- Multiple perspectives (for example, "standard" medicine and "alternative" medicine). This problem arises from our inability to incorporate all perspectives into one structure or scheme, no matter how richly articulated it is. We know that any one perspective limits what we see or learn, and diverse perspectives evolve as a result.
- Multiple situations/contexts
- Many different uses of knowledge

Implications

- One scheme or many?
- One representations vs. multiple representations
- Limitations on mapping between schemes

Role of classification in bridging diversity

Classification should honor diversity by reflecting different perspectives. But classification should also bridge diversity by mediating between different points of view, different knowledge and cultural systems. For example, a classification of concepts in "alternative" medicine could include scope notes and relationships that relate its concepts to concepts in "standard" medicine. By elaborating concepts, concept relationships, and conceptual structure in different realms, classification can help identify commonalities and differences and the nature of differences, supporting an effort at sharing and refining conceptual structures.

Theme 3. The quest for unity. Multi-purpose classifications, reuse

- Classifications require considerable intellectual investment, so one would like to reuse them. However, this practice causes tension with diversity.
- Classification modules that can be used in different schemes:
- How do we build modular ontologies to better represent dynamic domains? These would be ontologies that could flexibly extend the working ontology, for example extending the ontology of basic business processes by adding a module about auctions.
- How can we build classification schemes that store basic-level (mid-level) attributes that are neither too abstract nor overly specified so that they can be used effectively

by people in a variety of contexts, when we know neither who the people are nor what the contexts are?
- The mapping of ontologies one to another must include more than just terms and their relationships, but must also include information about the context/situation.
- Can a thesaurus be reorganized for multiple purposes?
- Is it possible to reorganize an existing thesaurus into a "navigational ontology" to support searching and browsing? Or does such a tool have to be created initially with these goals in mind (re: the previous question)? Can one thesaurus be reorganized in different ways to serve multiple purposes, such as searching, navigation, instruction, and "stimulation" (creativity)?

Theme 4. Types of knowledge covered in classifications

- Role and importance of all knowledge types
- Most classifications deal with (static) domain knowledge
- Additional approaches are needed to support users, such as
  - Problem schemas as organizing principle: A classification of problems by problem type, such as *fix a device* (*fix a car*, *fix a washing machine*), *buy something*, *write a computer program*, giving for each problem a schema that specifies aspects to be considered in solving the problem; information, people, material needed for solving the problem; procedural steps for solving the problem.
  - Functions as organizing principle, for example, technical components classified by all the functions they could serve
  - Classification of cases for case-based reasoning or for education and learning
- *Implications*: Importance of stepping back from what we "know" about building an ontology based on domain knowledge

Theme 5. Orientation of classification: Users' conceptual structures or intrinsic logic of the domain

Should classification reflect
- the users' conceptual structures;
- the intrinsic logic of the domain (as elaborated by the classifier) on which AI inferences could be based from which users could learn?

How can a classification be constructed that mediates between these two orientations?

Theme 6. Types of relationships in a thesaurus / classification / ontology

- Traditional thesauri use just BT/NT and RT as conceptual relationships
- Do we need a richer set of relationship types (as in SERUBA or the Vesalius ontology)? How are the relationships beyond the standard hierarchical relationships determined and how far can they be taken?
- In a visual environment, such as anatomical images, what other types of relationships besides those discussed in the Vesalius navigational ontology, could be developed? Are these limited by specific visual domains as well?
- How successful is the idea of a navigational ontology in a non-visual environment? This implementation draws upon the ideas of structure and function for navigation; text-based thesauri rely heavily on hierarchical relationships (structure) with function (related terms) being an unstructured grab bag, so to speak. Is it possible to transfer the idea of conceptual navigation incorporating both structure and function to a strictly text-based domain?
- In a non-visual environment, should the multiplicity of existing RT types be made explicit to the user? Making RT types explicit may enable people to recognize relationships that they may have otherwise omitted from their search. Related to this, to what extent are RTs bounded by a particular domain? And do specific types of RTs occur more frequently in a particular domain?
- What are the cultural issues between languages – are some of the relationships more apparent in some languages than in others? (The diversity theme)

Theme 7. Display and user interaction issues

- Classified display of search results is useful
- A wide range of methods for displaying classifications is available
- Should users interact
  - with the classification structure – concepts and their relationships;
  - with a categorized list of results; or
  - with a combination?
- Regarding the display of relationships among categories: would there be a benefit to users from displaying relationships among categories rather than just displaying category names? Would this add too much complexity? How should concept relationships be displayed (concept maps etc.)?

Theme 8. Practical issues

Classified displays are useful, *but*

- Constructing classifications manually is expensive
- Indexing items manually is expensive

What can be automated? (Session 3)

Dagobert Soergel
ds52@umail.umd.edu

With contributions from the session rapporteurs Edie Rasmussen, Corinne Jörgensen (extensive report on session 2), Linda Rudell-Betts, Jian Quin, and Barbara Kwasnik