Barbara H. Kwaśnik (Syracuse University)
Mary Grace Flaherty (Syracuse University)

Harmonizing professional and non-professional classifications for enhanced knowledge representation

Abstract
We compare two separate but related classification schemes in the area of medical information in order to better understand how they might be used together and inform one another. First we examine a “professional” scheme, the thesaurus of Medical Subject Headings (MeSH) produced by the U.S. National Library of Medicine. We contrast this with the “naïve” scheme used by the consumer health website, WebMD.com. Using the term autism we compared the strengths and limitations from the perspective of vocabulary, syntax and classificatory structure, context, and warrant. We conclude that in terms of vocabulary and concepts, MeSH may benefit from WebMD’s approach to ongoing updates and currency as well as the contextualization of terms. At the same time, WebMD may benefit from some form of vocabulary control for richer expansion of terms and archival retrieval.

1: Introduction
In the call for papers for this conference the organizers stated:

“Although many systems have been designed for the purposes of specific domains and communities, the new situation of interconnected global knowledge means that often we cannot anticipate which users will access our knowledge resources. This poses the question of how to deal with both global and local surroundings and needs […] with a diversity of traditions and special viewpoints […] within the framework of a global platform.”

Our aim was to systematically investigate the links and mapping of separate but related classification schemes in order to facilitate their use together, thereby extending our ability to use them as classifications are meant to be used: for representation, exploration, edification, and retrieval. We do not believe that schemes devised by different communities and for different purposes need to be kept in separate silos. Instead we believe they can inform and shape each other.

Most of us stand in several worlds, or at least we understand that different worlds construe the phenomena of interest differently. Thus an amateur gardener understands that there are other discourse communities that deal with the various phenomena of gardening: botany, chemistry, ecology, and so on. The wine lover knows there is a chemist out there studying the properties of fermentation. These scientific/professional communities overlap somewhat – depending on the intensity of the interest, but they also form separate domains in terms of what is “of interest.”

To the botanist a “weed” is another plant. Does he or she care that it wreaks havoc in gardens? Does the gardener care about the botanical attributes of the weed? Often they do care. In some ways these two discourse communities may talk to each other, but frequently it is only through the intermediation of someone who understands both worlds. Increasingly, though, we have the responsibility to understand more than just our local view. Consider for example the classification of “endangered plant species”. How does that classification overlap with the domains of scientists, economists, gardeners, developers, politicians, and ordinary citizens?
Previous work

Beghtol (2003) makes the distinction between classifications created for retrieving the literature of a field and those created for some other goal. The former she calls “professional” classifications and they emerge from a cyclical process in which scholarly activity leads to new knowledge; literary warrant becomes extensive; consensus among scholars develops; classification is created or revised to accommodate new knowledge; documents are classified in new scheme; and eventually the cycle repeats (p. 70). Such classifications rely on the norms and consensus of research communities, and are bounded by the need to provide the truest reflection of existing norms. In other words, the classification does not so much guide the production of new knowledge, but rather, reflects it and thus allows us to organize and tap that existing knowledge in a productive way. She goes on to distinguish such classifications from naïve classifications that are invented to serve a new need, and that evolve from a particular contextualized goal rather than from a commonly held one.

In an effort to understand the relationships among classifications, our previous work has focused on how classifications differ from each other structurally and culturally. For example, one study (Kwaśnik & Rubin 2003) examined how the kinship terms of 14 different languages mapped onto a set of kinship terms in English in the DDC and LCC. We looked for differences in the scope of each term. Did it cover the same entities? We also examined empty lexical or conceptual categories, differences in criteria for distinction, and most revealing, differences in how terms are used in practice. This comparative analysis allowed us to identify patterns of issues such as insufficient specificity to describe the phenomenon in both cultures, or culturally significant terms that may not map accurately to all languages.

In another study (Kwaśnik & Chun 2004), we explored how a translated classification (DDC in English to Korean Decimal Classification) exhibited the effects of cultural differences that would not necessarily map readily. In conducting the analysis we employed similar criteria as above and added those specific to bibliographic classifications. We identified misclassifications, where forcing a concept into the classification of another language may violate how it is construed. There are also differences in the extended uses of terms and ambiguities resulting from how terms are used in practice.

In this study we wanted to introduce another aspect – namely the goal of the classification as it is manifested in the structure and vocabulary. In his study of the Mayan Tzeltal folk zoology, Eugene Hunn (1977, 39-72) discusses the differences between folk and scientific classifications. The differences are not that his informants had different powers of observation or perception. Even so, their classification does not put humans in with animals, since only humans have a soul. The difference is what he calls “the intrusion of deductive criteria” (p. 64). In this case there is a Tzeltal story that explains how monkeys found themselves on the animal side of the classification, even though anyone can see they resemble humans. Thus no matter what the inductive evidence of similarity, it is not sufficient to outweigh the deductive mythology. In this sense, Hunn is echoing Beghtol’s distinction of naïve classifications being guided by particular goals and models of discovery. The classificationist imposes a structure and logic that suits the particular view.
3: Method
To demonstrate this principle we chose the topic of autism, a particularly active segment of two classifications in the medical field. One example was drawn from the “professional” scientifically vetted MeSH vocabulary, and the other from WebMD.com, a site meant for the general public.

Our plan was to perform a term-by-term structural and descriptive analysis of the underlying classification scheme, and at the same time to analyze for the “intrusion of deductive criteria”, such as the influence of guiding goals, users and uses. We wanted to identify crossovers and mutually informing merger points. Can one make some general recommendations of how to harmonize these two systems so that exploration in either one is enhanced given the differences in goals? We chose two contrasting systems supported by different classification schemes in order to explore the landscape of each with an eye to identifying how they might be utilized to leverage each other to enhance retrieval.

*MeSH.* The Medical Subject Headings is the controlled vocabulary thesaurus developed by the U.S. National Library of Medicine (NLM) for aiding in searching the biomedical literature; its online origins date back to the 1960s. MeSH uses a hierarchical “tree” structure, so that searching can be performed at varying levels of specificity. In the 2009 MeSH, there are 25,186 descriptors. There are also over 160,000 entry terms to aid in identifying the best preferred MeSH heading; for example, fibroid is an entry term for the MeSH heading leiomyoma. MeSH is used by the NLM for indexing articles for their PubMED/Medline database, which currently includes literature from over 5200 biomedical journals, and is primarily used by researchers and clinicians in the biomedical fields. MeSH is also used for cataloging books, documents, and audiovisual materials for the National Library of Medicine’s database.

*WebMD.com.* WebMD is a commercial website, operated by the WebMD Corporation. While the WebMD Corporation supplies information technology for health-care providers, it is most widely recognized for its consumer health information website (Funding Universe 2009). The primary audience is the general public; in July 2008 over 17 million visitors were reported to have visited the website (Reisinger 2008).

We performed searches in WebMD and Medline for autism. The original intent had been to do a term-by-term comparison, but this was not possible since the structure of the underlying vocabularies in both resources is not comparable at that level of specificity – or, put another way, such a comparison would not yield insights as comparing one bibliographic scheme to another might do, or one controlled vocabulary to another. Instead, our overall approach was to compare and contrast hubs of information centering on autism. Towards this end we mapped out the results of our searches and looked for clues for gaining insight into how each of these schemes might benefit from some of the strengths of the other.

4: Results
Starting with Medline, which uses the MeSH structured vocabulary as its framework, the search for autism yielded the following path:
A dozen or so entry terms, including autism itself, leads the searcher to this MeSH term. Autistic disorder is also coupled with approximately 30 possible subheadings including the following sample:

<table>
<thead>
<tr>
<th>--Complications</th>
<th>--Economics</th>
<th>--Prevention &amp; Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>--Chemically induced</td>
<td>--Epidemiology</td>
<td>--Psychology</td>
</tr>
<tr>
<td>--Diagnosis</td>
<td>--Etiology</td>
<td>--Rehabilitation</td>
</tr>
<tr>
<td>--Diet therapy</td>
<td>--Genetics</td>
<td>--Therapy</td>
</tr>
<tr>
<td>--Drug therapy</td>
<td>--Immunology</td>
<td>--Virology</td>
</tr>
</tbody>
</table>

These term combinations lead the searcher to a potentially rich array of resources related not only from the perspective of medicine, such as therapies and etiology, but also by other contextual aspects, such as psychology and economics. Thus, to summarize, the Medline search allows further searching for more general topics, such as childhood developmental disorders, and also among related and more specific topics as suggested by the subdivisions.

WebMD functions differently. There is a master index of over 6500 topics arranged alphabetically. These lead to 40 “Health Centers”, some with subdivisions and several other departments, all related to health information. In searching for a topic such as autism the master list directs the user by way of a shallow tree to “centers”, one of them being the center comprising a collection of resources of all kinds related to autism.

Once at the Autism Spectrum Disorders Health Center, the user finds a busy page of dynamic content: features, hot topics, menus, ads, and graphics that are regularly updated. It really does resemble walking into a “center” with the latest and most interesting resources highlighted in a variety of ways.

In terms of classificatory structure, what looks like the main menu, labeled Topic Overview includes the following subtopics:

- Symptoms
- Exams and tests
- Treatment overview
- Home treatment

While these are not formally subdivided, the first four are essays, written by the WebMD staff, summarizing that topic. Each essay contains further embedded and linked vocabulary. For example the essay on Treatment includes:

- Behavioral training and management (which includes references to):
  - Applied behavior analysis (ABA)
  - TEACCH (Treatment and education of autistic...)

WebMD Home
Brain & Nervous System Health Center
Autism Spectrum Disorders Health Center
Another list, labeled Topics Related to Autism includes an assortment of other links, including:

Asperger’s Syndrome  
Autism: Tips for caregivers  
Depression

Yet another “Hot List” features more essays that also include terms and links, e.g.:

What’s It Like to Have Autism  
Autism Exams & Tests  
Therapies for Autism  
5 Types of Autism...

Obviously there is no controlled vocabulary. Terms that seem related or synonymous appear in several lists and don’t always lead to the same material. It’s possible to meander around viewing similar but not identical “packages” of information, each, in turn, with its own set of terms and links.

5: Discussion
So, what is going on here? First a reminder: we purposely chose two contrasting schemes – one a “professional” classification and one “naïve”. MeSH is designed to index and search the biomedical literature. WebMD is designed to provide synthesized, short summaries written by the WebMD staff (not citations), that provide direct information, definitions, and further links. Each resource has formidable strengths, but also limitations.

Vocabulary. The MeSH vocabulary is meant to be stable and relatively permanent. It is vetted by experts, but also relies on other authoritative sources. In this case, autism is a term that has undergone change. WebMD uses the more current autism spectrum disorder, while MeSH still waits for the DSM IV to update its terms before updating it and making the change. Medline uses articles or resources as the starting points, which are then indexed with terms from MeSH. Using these terms assumes a certain level of sophistication, even though there are navigational aids that operate invisibly.

WebMD uses terms in the text that can yield interesting combinations that may not appear in MeSH. On the other hand, WebMD’s vocabulary is a bit of a mess. Once a resource has been found it’s difficult to repeat a search and remember which term yielded a given result, since they are not all integrated neatly. For example following links for treatments and therapies yields an overlapping set of results, but not always consistently. WebMD, in being opportunistic in its choice of terms, tends to use those that are the most widespread – that is, the ones that appear in the popular as well as scholarly press and will simply update it if that term changes. MeSH, relies on consensus before approving a term.

Syntax and classificatory structure. While the MeSH vocabulary has an elaborate structure that uses the consensus view of the health fields as the framework for the core relationships of the tree, the use of the extensive array of facet-like subdivisions allows many flexible combinations that make the system fairly expressive. WebMD, on the other hand does not seem to operate on any discernible structure. The terms and themes are linked as needed according to the judgment of the authors. This yields a natural flow
from one area to another, but not predictably or symmetrically. That is, each search is a new experience. A certain degree of experience with the site allows a person to understand what sorts of things might be found, but the classificatory structure (what little formal structure there is) is not how they are found.

**Context.** What struck us as the biggest difference between the two “schemes” was not the vocabulary or the differences in classificatory attributes, but that MeSH is structured but not contextualized. That is, the main tree headings and the subdivisions can be arranged as needed, but there is no ontological help in forming a bigger picture of a topic as complex and multifaceted as autism spectrum disorder. Each aspect of the topic has to be assembled afresh. This puts a great deal of control in the hands of the searcher, but also produces some fragmentation. When you visit the WebMD site, on the other hand, the information has been packaged in such a way that all aspects are easily browseable – starting from the medical aspects, but stretching out easily into social, psychological and behavioral aspects as well. The price for that, however, is that while the searcher is left to soak things up, a direct search is not very satisfying or complete, and one has no idea of whether it is exhaustive or comprehensive.

**Warrant.** Another big difference is that MeSH purports to be authoritative, based on scientific and literary warrant (Beghtol 1986), and thus forms a sound basis for communication and continuity. WebMD has a different philosophy, based on cultural or educational warrant, the presence or absence of concepts is based on what is in demand: “We pride ourselves in knowing our audience’s needs […] We know that there is a difference between using a health site for health ‘performance’ issues (e.g., flat abs) vs. health research needs (e.g., ‘What is type 2 diabetes?’) vs. community support […] Our mission is to fulfill all these needs in the most appropriate ways possible […] Rather than filtering certain types of information that may or may not be applicable to any one individual’s personal health, we rely on you, our reader, to choose the information that is most appropriate for you.” (WebMD 2009)

Thus, the deductive criteria that Hunn (1977) mentioned come into play here. While the authors of the content in WebMD certainly are aware of the MeSH view of autism as articulated in the medical literature, they are responding to an audience who construe the syndrome itself and all its enveloping experiences in the world from different perspectives. Thus, the WebMD structure is left loosely un-authoritative. For example, issues that are controversial, such as the linking of autism and vaccinations, are included in the structures, along with factual and experiential narratives. The system welcomes many different discourses.

**6: Conclusion**

The aim of this paper was not so much to point out the differences between a naïve and a professional classification, but rather to speculate on how the two can be harmonized or leveraged to enhance strengths and minimize limitations. Based on the one modest example described above, we have the following suggestions.

In terms of vocabulary and concepts, a formal tool such as MeSH might build in some ways to informally update and keep things current, even before the terms and structure are finalized. Harvesting current usage and concepts from a site such as WebMD and giving them a “temporary” or “pending” status might help bridge the gap between the scientific and lay perspectives. They have different goals, but just as the
gardener and the botanist may have different interests, there are places where their interests overlap.

Conversely, while WebMD’s freewheeling free-text approach is less work up-front, there is a built in limit to what can be usefully presented and then archived. There is a sense that WebMD is supposed to be fresh, current, and responsive, and yet, as the valuable collection gets larger, there might be some advantage to systematizing, or at least recording the links. As well, the notion that certain topics can be expanded systematically using a vocabulary such as MeSH as a guideline for broadening or narrowing or faceting a search, might be worth exploring.

Finally, the aspect of WebMD that makes it so appealing is the sense that in the “Autism Spectrum Disorders Health Center” someone has taken the trouble to gather all the relevant pieces of information in one place. While the relationships among the concepts are not laid out in a schedule, the fact that they are embedded in essays and include definitions, helps tie various aspects of autism and its effect on people’s lives into an understandable whole. Yes, the navigation is confusing, and yes, the terms lack consistency, but in the end, once you’ve wandered in and out of the various features in the center you have a lively sense of what is “out there” in terms of this topic. By contrast, MeSH leaves the links loosely coupled. It would be helpful to find a way of embedding MeSH terms in narratives, making the relationships explicit so that a picture of the whole could emerge. The system has within it the makings of such an ontology.

Understanding how a classificatory structure works is the first step in planning for how it might be harmonized with another “sister” classification. These are first steps in providing some guidelines for mapping of disparate classifications so that the integrity and goals of each is maintained, while the synergy of the two working together is maximized.

References
Beghtol C., 2003, Classification for information retrieval and classification for knowledge discovery: relationships between “professional” and “naïve” classifications, Knowledge organization, 30, n. 2, p. 64-73.

Web documents have been accessed 29 September 2009.