Discussion Forum ■

Integration and Beyond:

Panel Discussion

Panel members: WILLIAM W. STEAD, MD, RANDOLPH A. MILLER, MD, MARK A. MUSEN, MD, PHD, WILLIAM R. HERSH, MD

This is the edited transcript of a discussion, among the authors and audience, that followed the presentation that led to the paper ''Integration and Beyond: Linking Information from Disparate Sources and into Workflow,'' which appears on p. 135.

Mark Musen: Bill, when you presented the three generations of integration, the implication was that the third generation is at hand. It was all present tense. I think all of us agree that architectures that allow us to encapsulate knowledge and data in ways that permit reuse are quite exciting. But are we really in the present tense? Have we really achieved these kinds of architectures and, in particular, when you go to the vendor demonstrations, what do you see of this?

Bill Stead: That is a very interesting question, because I think the third generation is more in hand than the second. I think ''generation'' may be the wrong word, because it suggests that the third supplants the second. Instead, techniques from each of the generations coexist in equilibrium.

For example, the UMLS provides us with mapping between codes for the kind of things you order for a patient (diagnosis, tests, medications) and the literature. At Vanderbilt we use this mapping to let you ask, ''What are the references relevant to the things that have been ordered?'' So, to that degree, third generation exists.

It is the second generation that is really hard, because it requires regularization. There's a difference between the Vanderbilt VEGETABLE and the Columbia MED, in that the Columbia MED relates the source vocabularies of the various feed systems (third generation), whereas the Vanderbilt VEGETABLE tries to build an enterprise-wide source vocabulary that is then reflected back into the source systems, prealigning their vocabularies.

Back to your question, 3M is an example of a vendor that has pursued an architectural strategy.

Member of the audience: I think one of the toughest things we all have to deal with is updating our dictionaries. In the simplest cases, the name of an organism is changed and we just have to do the maintenance. It is tougher, when, as with *Citrobacter*, they do genetic studies and say, ''Oh, it's really six different organisms, not one.'' We have the human genome project coming very quickly. Even that is just the tip of the iceberg. We're not only going to see all the genes; we're then going to see clinical tests based on gene expression. Essentially, you'll be able to look at something on the order of 180,000 gene products and whether they're up or down regulated. How are we going to integrate such an incredible amount of data at a time when we're going to also be changing how we think about these processes? Classification and simple mapping are not going to work, because the lumpers and splitters are going to be arguing furiously on a daily basis.

Randy Miller: The problems you mentioned are clearly on the horizon and very important. But at a simple level, people are people and all of what you're talking about doesn't change how people will present to their primary care providers. At least that part of what exists will not get torn apart. I think what you're talking about is very rich, very vast information overlays on top of what we already have. We don't have to throw out what we have, we need to be ready to extend the linkages. How that will be done is an unanswered question that will result in multiple research grants.

Affiliations of the moderator and participants: Vanderbilt University, Nashville, Tennessee (WWS, RAM); Stanford University Medical School, Stanford, California (MAM); Oregon Health Sciences University, Portland, Oregon (WRH).

This panel discussion with audience participation was part of the Cornerstone on Integrating Information, one of four Cornerstone sessions included in the program of the AMIA Annual Fall Symposium, Washington, DC, November 6–10, 1999.

Correspondence and reprints: William W. Stead, MD, Associate Vice Chancellor for Health Affairs, Vanderbilt University, 416 Eskind Biomedical Library, 2209 Garland Avenue, Nashville, TN 37232-8340; e-mail: ⟨bill.stead@mcmail.vanderbilt.edu⟩.

Received for publication: 11/1/99; accepted for publication: 11/18/99.

Bill Hersh: I think you allude to one of the key points, which is structuring the metadata with the right levels of granularity. Clearly, when we find an organism that can't fit in the existing framework, then that's a problem. But if we find that an organism just represents a subcategory of others, and if there's a good hierarchic structure, it can be fit in. The same goes, for example, for diabetes. People classify diabetes with this complication and that complication, but often we just want to know whether the patient has diabetes. Again, a good hierarchic metadata structure can overcome some of those problems.

I think we also need to recognize some of the practical limitations that face us. There are limits to the accuracy of the information that's in medical records; there are limits to the consistency in which people apply vocabulary terms. Computers can be completely precise in terms of mapping from this to that, but people will continue to have different conceptions of what a "grade II systolic murmur" is.

Bill Stead: I agree with both answers, but I want to continue to clarify what we are talking about. We get in trouble because people use words to reduce concepts to something that we can manage in our heads. So we lump, and person A lumps differently from person B. So we are each a "legacy system," and our information resources have grown from this starting point. I think we need to work at two ends of the spectrum. Whenever possible, capture data according to granular definitions. If we have an organism and we discover that it splits into six organisms, that's actually a very easy problem to solve, as you said. What you've got to do is say, "A is now B, C, and D and it mapped here." That is straightforward. That's the end of the spectrum where we can stay granular. For example, never store a doctor and the doctor's service as one piece of information. At the other end of the spectrum, where the granular definitions are not obvious, do not try to classify the data. Instead, tag a "clump" of information with metadata. This tagging, together with increasingly sophisticated extraction techniques, will be used to approximate meaning. Over time, we will get to a complete set of coded data by working from the two ends.

Mark Musen: I'm not sure that everything will ever be completely coded. Given the fact that the world is continuously changing, I don't think we can assume that Aristotle was correct that eventually there will be a classification that we will all accept. For example, I do not know whether gastric ulcer is an infectious disease or a gastrointestinal disease, and maybe it is both. As we continue to learn more about medicine and as our organizations change out from under us, I think we're going to be in the situation where the way we categorize the world is going to change. This is very hard stuff. Instead of working on the ultimate classification that will have all of the problems of the International Classification of Diseases, we need to build structures that not only allow us to enumerate the kinds of data that our programs operate on, but attempt as best as we can to enumerate the assumptions that we're making about our data and about the world. Then, as things change, we can, as human beings, try to update our ontologies. I think we have to be able to deal with changing worlds and with the fact that people and computers each need different views on the data, and that means different assumptions as well.

Bill Hersh: To reiterate Mark's point, some people have heard this quote, that "perfect is the enemy of good." We, especially us academic types, strive for perfection, but in reality the world is not perfect, and I don't know that everything will be perfectly coded. But we can reach compromises, such that we can code bits of information that enable us to do useful things.

Bill Stead: I think human beings are each different, but we have an underlying genetic code that we are in the process of discovering. Next, we are going to have to work out the problem of going from genotype to phenotype. When I say that I think in the end things will be coded, I think we're going to discover something that is to information what DNA is to people. It will be a very granular base set of building blocks, which will be rolled up into concepts much as genes produce proteins. So I do not want to go to one ontology or one classification. Still, I like having ontologies, particularly ones that clearly represent the difference between themselves and the others.

Member of the audience: I'd like to ask a question about capturing ontologies from multiple people. Imagine for a moment that knowledge freezes long enough for us to try to catch it. Do you have a vision of a tool that will allow multiple knowledge-domain people to act at once? To work out discrepancies in their visions?

Mark Musen: Put differently, the question was how do we deal with the fact that there is no overarching ontology? How do we build the tools that will allow us to try to achieve consensus in ontologies? I think the answer to that question is that we do not know. I'm being a little bit facetious, but philosophers have been trying to deal with that problem for 2,000 to 3,000 years.

I think you see two different approaches in the computer science community. You see the approach that Doug Lenat has taken. He is trying to create an ontology that he believes will provide all the knowledge that one needs to read the Encyclopaedia Britannica. Such an overarching ontology would need to capture most of human existence. The real problem, though, is how you ever validate the distinctions made in that ontology and have confidence that things have been captured in a way that is consistent and understandable? How do you record all the assumptions that you make while constructing the ontology? When you have concepts like "semi-tangible object" and "semi-intangible object," it's very hard to know for sure whether what one records about those distinctions really makes sense.

At the other end of the spectrum, you see people who really want a thousand flowers to bloom and who are not trying to achieve that kind of perfect alignment among views of the world. For example, the Knowledge Systems Laboratory at Stanford is trying to make constrained ontologies that deal with very narrow domains, so that the kinds of problems that you allude to do not happen, because the number of concepts in the ontology is relatively small. The answer lies somewhere between Doug Lenat's view of the world, that all we have to do is work hard enough and everything will fall into place, and the view that we can't possibly do this, so we have to have just a small number of constrained ontologies. We need to elucidate a set of principles that will provide the basis for tools that will help us try to, if not merge small ontologies, at least create the kinds of alignments that will allow us to bring them together in ways that make them useful.

Randy Miller: One of the things that I learned from my mentor, Jack Myers, is that as an informatician, as opposed to a philosopher or a computer scientist, you do not need to represent everything. If you have a problem at hand, you represent it at a level that is tractable and doable. If you do what Doug Lenat's doing, you can spend your entire career representing stuff that is not ever going to be used in a real system, because there is no way to apply it. While that may sound harsh, the reality is that we do not know how to represent time, severity of finding, and severity of illness well at all, but we can still build systems that do diagnosis or a good job of making recommendations for therapy. So you do not have to capture the world in all its infinite detail. The trick is to understand what the critical information is and represent things at that level. Otherwise, you get mired in detail.

Mark Musen: Let me underscore your last point. Doug Lenat actually felt pretty confident that his ontology covered all the areas that one would want to deal with, until last year, when HotBot contracted to use CYC as the basis for indexing Web pages. This contract showed, first of all, that ontologies have incredible commercial potential, but it also pointed out to Doug Lenat that there was a whole realm of human experience that was not well represented in the ontology. Specifically, there was a need to categorize different kinds of pornography, which Lenat had not thought about previously.

Member of the audience: Health Level Seven's development of a set of reference information models is one of the major efforts for creating a structure for ontologies in the United States. Can you talk about how your organizations are participating in the development of that reference information model (RIM) and how you are using your academic experiences to contribute to that effort among providers, academics, and vendors?

Bill Stead: Vanderbilt is an institutional member and a strong advocate of HL7. The central core of our communication subsystem uses HL7, and we build middle ware as needed to bridge between the core and legacy products. We have not put direct energy into the process for defining the reference information model. We use the HL7 model as a starting point, but we extend it as needed. In this way we incorporate it into immediate solutions to real problems, while providing useful information about future directions.

Bill Hersh: None of us has been involved directly in that effort. However, our research into the nature of ontologies and the vocabulary projects such as the Cannon Grouping should useful to the effort.

Mark Musen: I will just add that I think the vendor community is in the best position to work on ontology content, because they have the most direct connection with the needs of end users. I think that academicians need to follow this work very carefully. We are, we hope, in the best position to be developing the kinds of tools that will help us examine ontologies, relate them to each other, and allow them to evolve as our understanding of the world changes.

Randy Miller: I have a slightly contrary view, partly out of ignorance about HL7 RIM. The key question is what problems it is trying to solve. That should drive what the content is. If you can state the problems it is going to be used to solve, then you can say whether it should clinically rich. In that case it will require lots of input from academic clinicians. If it is to solve the problem of interchange of data among vendors, then it needs vendor input. But until you explicitly state what it's going to be used for, just building it for the sake of building it is not useful. I know that the HL7 RIM is not being built that way. I am just saying that I think that's the way to address your question, to seek the specific purpose before giving an answer.