125TH MEETING OF THE NATIONAL CANCER ADVISORY BOARD (NCAB)

MEETING OF THE AD HOC SUBCOMMITTEE ON BIOINFORMATICS VOCABULARY

February 11, 2003

3:10 p.m. – 4:10 p.m.

Subcommittee Charge

Dr. Chen called the Subcommittee Meeting to order and stated the charge to the Subcommittee. He reminded the Board that the need for this Subcommittee was identified at the December NCAB meeting and that this was its first meeting. Dr. Chen asked Dr. Kalt to amplify on the Subcommittee's charge.

Dr. Kalt indicated that harmonizing informatics within the Institute, between NCI and its awardees, as well as at national and international levels is a critical issue. The NCI NCAB is uniquely qualified to begin considering issues associated with trying to connect the increasingly expanding circle of databases to extract information both to manage programs and to create new knowledge. A critically important step, apart from software and hardware technologies, is the vocabulary that people use to talk and otherwise communicate with each other. Researchers, clinicians, and others must understand and use the same words for data so that everyone is speaking the same language and can share data. That is the issue that this Subcommittee is being asked to consider.

Status of Where We Are at NCI

Presentation by Dr. Francis W. Hartel, Ph.D., Director, Enterprise Vocabulary Services
NCI Center for Bioinformatics

NCI initiated the Enterprise Vocabulary System (EVS), a partnership between the NCI Center for Bioinformatics (NCICB) and the NCI Office of Communications (OC), which has become a sizeable effort to address NCI's needs for controlled vocabulary. EVS engages in research into theory and methods for organizing vocabulary, development of vocabulary content and services usable by both people and computers, and operation of servers and software systems to make vocabulary accessible to software systems as well as individuals.

There are several reasons why EVS develops vocabulary as opposed to simply adopting existing vocabulary. Most existing biomedical vocabularies are not computer-interpretable; they provide incomplete coverage of basic biology, especially genomics and proteomics; and inadequate coverage of cancer prevention and cancer specific clinical treatment and pharmacology, or provide inadequate linkage among these areas. Most are updated too infrequently to match the pace of cancer research. Finally, most of the existing vocabularies provide weak or no semantic content, which is vital to supporting medical informatics and bioinformatics.

NCI EVS provides two major vocabulary services. The first is the NCI Metathesaurus, based on NLM's Unified Medical Language System, with extensions and modifications to tailor it for cancer-centric uses. The second is the NCI Thesaurus, built by EVS, now with more than 25,000 concepts and 60,000 terms. A critical feature of this latter thesaurus is that machines can understand it as well as people. It represents semantic relationships among many concepts used in cancer science, science management, and clinical care. Its primary uses are: (1) to provide the computational linguistic basis supporting advanced information tagging and interpretation needed to meet NCI's requirements, and (2) to enable software to interpret automatically the semantic content of information.

The EVS has become vital to NCI's infrastructure. It is the base technology on which caCORE is built[1], which supports many NCI projects. EVS has produced state-of-the-art taxonomy of neoplastic and pre-cancerous diseases, the basis of the ongoing re-implementation of PDQ and support for significant improvements for NCI's portal website, *cancer.gov* and the Cancer Information Service (CIS). It also provides mapping services needed by CTEP in development of new common toxicity criteria and for reporting adverse events. The Vocabulary Executive Group, an advisory body composed of representatives of every NCI component meets regularly to convey the needs of NCI intramural researchers and the extramural and administrative communities, and to review EVS plans and priorities. The EVS staff also participates in national standards setting activities (ANSI/HISB and HL7), collaborate in interagency activities with the VA, NLM and FDA, and participate in the SNOMED Advisory Board. The EVS also requests reviews of NCI's major vocabulary developments by external experts.

Today, vocabulary is understood to be a highly dynamic and active resource. It is a critical component of HL7, the national standard for representation and exchange of medical information and underpins such research areas as proteomics, genomics, pharmacogenomics, and translational research, all of which depend on powerful vocabulary resources.

What are the responsibilities of this Subcommittee?

Dr. Chen discussed that while the EVS project is a leader in bioinformatics and computational semantics, both within NCI and within the larger bioinformatics community, it still needs guidance on its future scope of activities and help in formulating future development strategies for issues such as:

- Vocabulary and its relationship with standards development organizations
- Continuing research in computational linguistics and natural language processing
- How to best establish working relationships with the larger cancer community to make them aware of what the NCI EVS project is doing and to help them leverage these activities with other public and private entities
- Interfacing on important technical, societal, and scientific trends affecting ongoing development of this resource.

Dr. Chen suggested a possible approach be to form ad hoc working groups, each group to focus on one important topic or issue, members of the groups to be selected for their knowledge of the topic, and the group to prepare a report for review and approval by the Subcommittee. Dr. Chen also showed several questions that the Subcommittee might consider. He then opened the Subcommittee for discussion.

Dr. Love asked about the lifetime of the EVS' NCI Thesaurus, since it was indicated that older thesauruses had limited value today. Dr. Hartel indicated that new methodologies being used, such as description logic, make the current thesaurus logically consistent and machine interpretable and therefore the basic structure will not go out of date. In addition, EVS continues to do information science research to develop better, more expressive tools. However, he warned that once one gets involved in the vocabulary building business, one has to stay in the business to remain current. Also, other collaborating organizations, such as the Veterans Administration are also using description logic, so the entire field is growing together.

Dr. Prendergast asked whether the objective of the EVS project is to develop a single unified vocabulary that people would migrate to using. Dr. Hartel indicted this was not the objective, but rather to develop a vocabulary structure that will satisfy current and future needs of NCI and provide practical utility to support endeavors such as translational research. A single global vocabulary structure for oncologies would not succeed because one can't define the lowest level base foundation on which to build. Also, proliferation of terms is a never-ending problem. Dr. Prendergast than asked how we can define a

[1] CaCORE is a package of object models, databases, controlled vocabularies, and Application Programming Interfaces (APIs) for genomic and clinical application development.

vocabulary that will unify how disease is described, treatment described, etc. Dr. Hartel indicated that the HL7 standards work is attempting to standardize messaging, but allows for representations for each national domain—using each country's nomenclature with a mapping algorithm. Dr. Hartel also discussed the value of partitioning representation of needed information among object models, metadata models and vocabulary models as a way to gain leverage on the problem raised by Dr. Prendergast. The Board than asked what this Subcommittee and Working Groups should be addressing.

Dr. Chen showed a slide with several questions that need to be addressed, such as:

- ♦ Should NCI seek to become the focal point or representative of the cancer community in international and national informatics standards development, i.e., how broad should the EVS project scope be?

Several other questions were voiced by Board members—how do you get compliance with the vocabulary? How do you engage the global oncology communities? Dr. Norton suggested that before the Board or Subcommittee attempts to provide advice to the EVS, they need other pieces of the puzzle. He indicated that the Internet has made compatibility of language essential and that newspaper reporters need a common vocabulary. Therefore, the Board needs brought up to date not only on what EVS is doing, but what other competing systems are already in existence and what others are currently under development. He also reiterated that there is substantial impact of this work on such programs as the P30/P50, and that decisions on the scope of EVS could have substantial budget and workforce resource impacts.

Dr. Chen replied that the Subcommittee's intent is to develop working groups to address just these types of issues. Dr. Love indicated that NCI is a natural place for this work to be done, but asked why the Subcommittee can't take responsibility, rather than forming working groups. Dr. Chen indicated that this would be acceptable, if the Subcommittee could bring in needed additional expertise to address technical issues. The Subcommittee can then start with a review of what is already in place and being done in other Federal agencies, other global organizations and in private organizations. Dr. von Eschenbach suggested that one place to look is the intelligence community, which is struggling with this same type of problem, collecting information from many sources and trying to integrate and interpret it, turning it into knowledge. The Board also indicated that it would like to hear about efforts going on beyond the NCI to see if there are resources that could be adapted and applied to the EVS project. This effort should also attempt to find opportunities to partner and synergize and share costs and resources in getting this accomplished.

Conclusion

Dr. Kalt summarized the discussion by stating that this is a start and it is obvious that the Board thinks this is an important issue and that it is appropriate to have a Subcommittee to address terminology and to address issues surrounding bioinformatics. The second suggestion is to make an agenda item to get additional background information on where the field of bioinformatics vocabulary stands at the NCI, in a more global context. Another issue to consider is whether to bring into the discussion an extramural perspective for the Board to get a better sense of the issues and the structure internally. Since there was no further discussion, Dr. Kalt asked Dr. Chen to take this as an action for the Subcommittee. Dr. Kalt than reminded the Board of the members who volunteered for the Subcommittee to include:

Dr. Moon Shao-Chuang Chen, Chairperson	Dr. Susan Love	Dr. Amelie Ramerez
Dr. Elmer Huerta	Mr. Steven Duffy	Dr. Kenneth Cowan
Dr. Franklyn Prendergast	Ms. Lydia Ryan	Dr. Ralph Freedman

Dr. Love asked how this Subcommittee was to meet. Dr. Niederhuber suggested that the initial meeting take place by conference call. Dr. Kalt reminded the Chair, Dr. Chen, that a Subcommittee was empowered to bring in experts to assist in its deliberations and discussions.

The Subcommittee meeting was then adjourned by Dr. Niederhuber.

Moon S. Chen Jr. 2/12/03

Dr. Moon S.-C. Chen, Jr. Date
Chair

Francis W Hartel 2/12/03

Dr. Frank Hartel Date
Executive Secretary