Creating a Metadata Repository in Support of Clinical Research

Joyce C. Niland, Ph.D.
City of Hope National Medical Center
Department of Biostatistics
1500 East Duarte Road
Duarte, CA 91010 USA
Jniland@coh.org

1. Introduction

A metadata repository is a key component to a viable clinical research data system, one that is often overlooked and underappreciated by developers and users of such systems. This paper describes the process we've undertaken at City of Hope National Medical Center to develop and put in practice such a repository.

2. Background

Over 12 years ago the Biostatistics Department of the City of Hope National Medical Center in Duarte, California created a centralized networked clinical trials database system. The Biostatistics Information Tracking System (BITS) contains data on over 1000 clinical trials conducted in support of finding the causes, prevention and cure of cancer. BITS was originally developed using the Advanced Revelation software system, deployed over a PC network. The system is password protected to restrict access to research staff and investigators. BITS incorporates all data required for protocol administration, cancer center reporting, and survival analysis, and employs electronic interfaces to demographic, laboratory data, and HLA data.

For all patients registered onto a clinical trial, data on eligibility status, informed consent date, diagnosis, study arm, dates on and off treatment, follow-up interval, last contact date, relapse and survival status are computerized in BITS. Protocol-specific data include the prior treatment, protocol treatment, toxicity, treatment response, and any other results required for the final analysis. Over 100 menu-driven reports are available in BITS for National Cancer Institute (NCI) reporting, tracking of accruals, operational reports, and simple study work-ups or analyses. At the time of a full study analysis the data are exported from BITS into SAS for programming by the department's biostatisticians. The system now is being migrated to MS SQL Server 7.0, with Web-based screens and scannable forms as the new interface applications.

3. Creation of the Metadata Repository

A critical lesson learned during the construction and subsequent development of BITS is that the *metadata* are key to a sound, effective data system. Metadata are 'data about the data', that is information regarding the type and meaning of the data that stored in an electronic system. The metadata consist of two components, the 'business directory' that provides a definition of each data element, along with the key words, synonyms, and directives for collecting each element (Gray and Watson, 1998). The business directory information is supplied by the users or the 'steward' of the database system. The technical directory includes information obtained directly from the data model itself, including the field format, length, label, table storage, target reports, etc. Often these data can be imported directly from the underlying database model of the system.

At City of Hope we have created the electronic storage system for this information in the form of a metadata repository. The steps in planning our metadata repository have included: writing of the project specification and gathering user requirements; analysis of data dictionary requirements; creation of the metadata repository model; validation of this model; construction of the physical database; conducting an export of existing metadata from BITS into the metadata repository; and updating or inserting additional business and technical data as need.

The next step was to create the design specifications for the Web application front end to edit and access the data stored in the metadata repository. Our Web applications are developed using Active Server Pages (ASP) and Javascript. An analysis of the structured data available in the metadata repository was carried out. Content experts were called upon review the existing definitions and user directives, and complete the population of the business portion of the metadata repository by writing or adding new components as needed.

A critical step which is often overlooked is the creation and routine implementation of Standard Operating Procedures (SOPs) for on-going timely maintenance of the metadata repository. Additionally it is crucial to write an appropriate user manual, and provide training to the Operations Specialist charged with keeping the metadata current and accurate at all times.

After populating the metadata repository with the BITS metadata elements, we desired to utilize this system to harmonize and align data elements across two additional data systems that have been developed by and/or are maintained within the Department of Biostatistics. Therefore the metadata repository was augmented to contain additional metadata elements for our national outcomes research data system, and the data system for California State reporting of all cancer cases. The former is an Internet-based data system created to collect, manage, quality control, and analyze data for a national outcomes research study among the nation's leading cancer centers (Niland, 1998). The California State-mandated data system is required for the aggregation and reporting of cancer cases to facilitate future epidemiological research. After loading the additional metadata from these two systems it was necessary to analyze the overlap and discrepancies among the data elements stemming from the disparate origins, and to discover any existing commonalities among these data elements. The classification and synonym coding components of the metadata repository greatly facilitate such vocabulary alignment/discovery tasks (van Bemmel and Musen, 1997).

4. Future Plans

The reporting tools of the metadata repository are being designed to allow printing of independent data dictionaries from any one of the represented data systems, a master data dictionary for all systems, or subcomponent dictionaries within a system, such as outcomes research by disease under study. This system also will facilitate the training and quality assurance of data being collected for statistical analysis. Our ultimate goal is to expand and utilize this metadata repository approach to manage data contained in a research data warehouse (Bell, 2001). This will represent the optimal means to facilitate future complex, multi-disciplinary analyses, such as genotype-phenotype correlations (Rindfleisch and Brutlag, 1998).

REFERENCES:

Gray P, Watson H. Decision Support in the Data Warehouse. Prentice Hall, pp. 62-65 (1998).

Niland J. NCCN Internet-Based Data System for the Conduct of Outcomes Research. *Oncology*, 12:11A (1998).

Van Bemmel JH, Musen MA. Handbook of Medical Informatics. Springer, pp 88-89 (1997).

Bell L. Metabusiness meta data for the masses: administering knowledge sharing for your data warehouse. Journal of Data Warehousing, 6:2 (2001).

Rindfleisch TG and Brutlag DL. Directions for clinical research and genomic research into the next decade: implications for informatics. *Journal of the American Medical Informatics Association*, 1998;5:404-411.

RESUME:

This paper describes the creation of a metadata repository to store documentation regarding data contained in clinical research systems at City of Hope National Medical Center. Such a repository is critical to the efficient management and integration of data systems in support of biomedical research.