The Integrated Data Repository: A Non-Traditional Data Warehousing Approach for the Translational Investigator

Marco J. Casale[3], MS, Prakash Lakshminarayanan[1], MBA, Mark G. Weiner[2], MD, David A. Krusch[3], MD, Russ J. Cucina[1], MD, MS, Rob Wynden[1], BSCS
[1]University of California, San Francisco, CA; [2]University of Pennsylvania, Philadelphia, PA; [3]University of Rochester, Rochester, NY

Abstract
An integrated data repository (IDR) containing aggregations of clinical, biomedical, economic, administrative, and public health data are key components of an overall translational research infrastructure. However, most available data repositories are designed using standard data warehouse architecture using a predefined data model, which does not facilitate many types of health research. In response to these shortcomings we have designed a schema and a just-in-time ontology mapping service (JIT-OMS)[1] which will facilitate the creation of an IDR by directly addressing the urgent need for terminology and ontology mapping in biomedical and translational sciences and give biomedical researchers the required tools to streamline and optimize their research. The system will dramatically lower the barrier to IDR development at biomedical research institutions to support biomedical and translational research, and will furthermore promote inter-institute data sharing and research collaboration.

Introduction
Most repositories are designed using standard data warehouse architecture, with a predefined data model incorporated into the database schema. The traditional approach to data warehouse construction is to heavily reorganize and frequently modify source data in an attempt to represent that information within a single database schema. This information technology perspective on data warehouse design is not well suited for the construction of data warehouses to support translational biomedical science.

The purpose of this poster is to discuss the components which would facilitate the creation of an IDR by directly addressing the need for terminology and ontology mapping in biomedical and translational sciences and the novel approach to data warehousing design. Further, a derivative outcome of this work is the inter-institutional exchange of ontology maps which can benefit translational research.

Discussion
There are several challenges posed by IDR projects geared toward biomedical research which do not apply to the construction of most commercial warehouse implementations: 1) integrity of source data Regulatory requirements and researchers demand clear visibility to the source data in its native format to verify it has not been altered; 2) high variability in source schema designs - IDRs import data from a very large set of unique software environments, from multiple institutions, each with its own unique schema; 3) limited resources for the data governance of standardization - widespread agreement on the interpretation, mapping and standardization of source data that has been encoded using many different ontologies over a long period of time may be infeasible. 4) limited availability of software engineering staff with specialized skill sets.

We have developed an alternative approach that incorporates the use of expert systems technologies to provide researchers with data models based on their own preferences, including the ability to select a preferred coding/terminology standard if so desired.[2] We believe that such an approach will be more consistent with typical research methodologies, and that it will allow investigators to handle the raw data of the repository with the degrees of freedom to which they are accustomed..

References

1. Wynden R, et al.. *The Integrated Data Repository: Ontology Mapping and Data Discovery for the Translational Investigator* Proc: AMIA Summit 2009

2. Noy NF et al. Protégé-2000: An Open-Source Ontology-Development and Knowledge Acquisition Environment. Proc. AMIA Symp. 2003; 953.