Developing Metadata to Organize Public Health Datasets

Michael D. Matters, PhD, MPH [1,2], Akaki Lekiachvili MD, MBA [3], Thomas Savel, MD [1, 4], Zhi-Jie Zheng, PhD [2]

[1] Public Health Informatics Fellowship Program, OWCD,CDC, Atlanta, GA, USA

[2] Division of Adult and Community Health, NCCDPHP, CDC, Atlanta, GA, USA

[3] Office of Informatics and Information Resources Management, NCCDPHP, CDC, Atlanta, GA, USA

[4] Office of the Director, NCBDDD, CDC, Atlanta, GA, USA

Abstract

The Centers for Disease Control and Prevention (CDC) has available a large number of datasets from previous and current surveillance and research.[1] Until now, these datasets have not been catalogued. Metadata would organize these datasets and enhance CDC's ability to efficiently use this data to quickly gain the broader view of the nation's health status to effectively carry out public health activities. This project was to develop metadata for cataloguing CDC datasets and a system that would allow researchers to search at least 95% of databases within CDC based on the most relevant criteria for research. It also explored the need to involve stakeholders and users in the project. The resulting metadata and system are available only to CDC researchers on the CDC intranet.

Background

Researchers at CDC have access to numerous sources of data. However, these researchers may not be aware of, or know the content of, other important datasets. The project's purpose was to create metadata for dataset documentation. The goal was not to develop comprehensive descriptive data, which might be too complex to be useful. The intent was to identify a subset of attributes that will be relevant and specific for CDC researchers and help them understand current public health issues and implement appropriate public health response. These metadata will enable researchers to search and catalogue available datasets and to streamline access through provision of relevant information.

Methods

The development process was begun by looking at existing metadata standards and similar initiatives in other content areas. Among these were the Census, the Federal Geographic Data Committee, the Dublin Core Metadata Initiative, the International Organization for Standardization, and the National Information Standards Organization. Relevant existing metadata was identified to develop a preliminary draft document. This led to the creation of an overarching list that was relevant and applicable to majority of datasets used at CDC. Joint action development (JAD) sessions were conducted with different stakeholders within CDC to refine the system. The sessions focused on metadata identification and definition of system scope and requirements based on stakeholder needs. The requirements for the prototype currently are being created. The most important part of the development of any system was stakeholder and SME (small and medium-sized enterprise) involvement; developing the web-based application will be relatively easy and secondary. This project also examined the importance of stakeholder involvement in the process and balancing the output between the different needs and requirements at the local and national level. The final step is to place these metadata on the CDC intranet where they are available to researchers.

Results

A comprehensive list of metadata and related definitions was created. Based on feedback during JAD sessions, the stakeholders were very receptive since they see that the metadata are developed around their needs. This first phase appears to be the most important predictor of success. Nevertheless, the final success will be known only after the system has been fully implemented. The system was designed to be created for and maintained by CDC researchers. As the system's stakeholders and users, they decided on most relevant attributes for the metadata. Since this effort was internal to CDC, it had to balance between national initiatives and local needs. However as we move the project forward, we are working to have these standards formalized through the Health Level 7 consortium and are exploring methods for wider applicability.

Conclusions

The Metadata Documentation for CDC Datasets provides a structured way to catalogue and search data. It balances between size and complexity to facilitate easy implementation. It also provides the ability to utilize a program-centric approach by implementation at a subset level and allows a certain level of customization.

References

1. Chute C, and Koo D. Public Health, Data Standards, and Vocabulary: Crucial Infrastructure for Reliable Public Health Surveillance. Journal of Public Health Management and Practice. May, 2003:8:3.