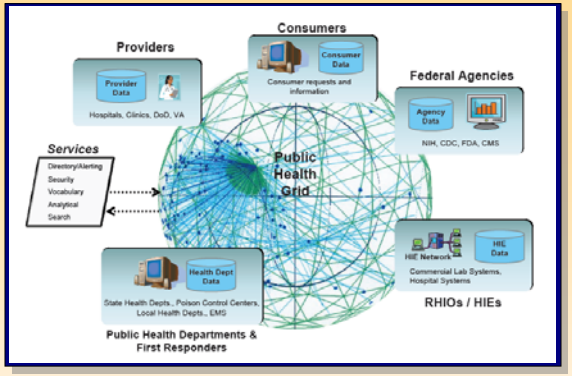
Capturing Metadata for the BioSense Surveillance System: Preparing for a Public Health Grid

Gail E. Scogin, MS, John Lindsey, Russell Gann, MS

Centers for Disease Control, National Center for Public Health Informatics

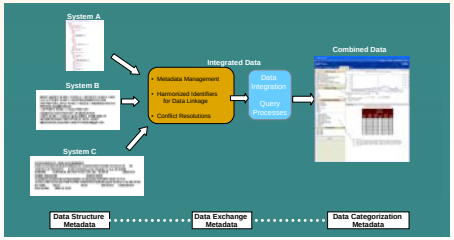
The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

The Public Health Grid is a vision for the future of Public Health information in the US.



- The technology to access data stored remotely is available, stable, and (relatively) easy to provide.
- However, querying and combining multiple, different data models (and structures) can be very difficult, requiring:
  - Metadata Management: Metadata is needed to fully describe the attributes, business context and processing of data.
  - Harmonized identifiers: Different systems may use different identifiers for the same data.
  - Naming resolution: Data elements may have different names in different sources, yet mean the same thing.

Semantic interoperability is a major design challenge in building an Information Grid.



- Comprehensive metadata management combined with the development and implementation of standards and governance will assist public health agencies to create and use service-oriented environments as the basis for data grids.
- However, service-oriented architectures present both opportunities and challenges for data sharing.
- Effective data sharing requires semantic interoperability so that data can be understood unambiguously and processed in a meaningful way.

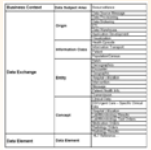
The Federal Enterprise Architecture Data Reference Model provides an initial framework for enabling data discovery and sharing.

- In 2002 as a part of the eGov initiative, federal agencies began using Federal Enterprise Architecture (FEA) reference models to classify and organize federal IT assets.
- The Data Reference Model (DRM) is intended to support the integration of data in federal IT systems. However, it can also provide a type of governance model for managing data in a grid environment.
- The DRM provides a standard means to describe, categorize, and share data. These standards are reflected within each of three areas:
  - Data Description: Provides a means to uniformly describe data, thereby supporting its discovery and sharing
  - Data Context: Facilitates discovery of data through an approach to categorize data according to taxonomies; enables the definition of authoritative data assets within a community of interest
  - Data Exchange: Supports the access and exchange of data where access consists of ad-hoc requests (such as a query of a data asset), and exchange consists of fixed, re-occurring transactions between parties

BioSense metadata has been captured based, in part, on the FEA Data Reference Model framework.

Using a DRM schema, CDC began capturing metadata associated with the processing of HL7 messages from participating hospitals and public health surveillance systems.

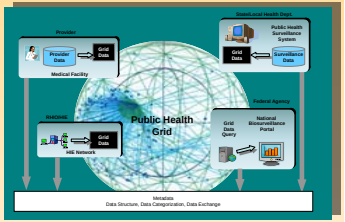
- The metadata is organized according to logical, business-related design elements which include data concepts, information classes, and entities.
- The captured metadata also includes more detailed structural elements such as the physical databases, tables, views and control structures in which data elements are stored.
- A BioSense metadata repository will provide a consistent and reliable means of access to information about BioSense data.
- Analysis and management of these metadata will support improvements in data architecture and governance as well as provide a resource for the discovery of data reuse opportunities.



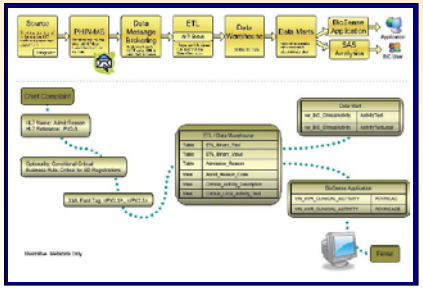
Illustrative Metadata Only

Comprehensive metadata management is key to semantic interoperability.

- To effectively query and integrate across multiple data sources requires a deep understanding of the different data models involved.
- Comprehensive metadata management includes metadata specified in the Data Reference Model:
  - Data Structure: Information about the data structures themselves
  - Data Categorization: Information about the business context of data within the data structures
  - Data Exchange: Information about the exchange and transformation of data between data structures
- In many cases, data sources present their own unique formats, requiring syntactic translation and semantic interpretation.
- Finding and querying data from multiple, different sources across the Grid will be complex. It will require an understanding of data content, context, format, and structure to accommodate the varying levels of conformance to available standards.
- Comprehensive metadata will be required for each data source to enable the Grid and resolve integration issues.



BioSense metadata has been expanded to include end-to-end data processing.



- A primary goal is the collection and synthesis of end-to-end metadata for data process flow, transformation, and visualization.
- A related goal is the creation of a metadata repository that serves as a searchable system to facilitate analysis and sharing of the data's business context, uses, quality, and potential suitability for other purposes.
- This repository can also help assess the impact of change, facilitate architectural improvements, and provide a stable source of information to support the design of data-centric reusable web services.
- Such a repository ultimately supports the integration of biosurveillance and other public health data.

Acknowledgements

- Vernicia Reese, Gideon Slifkin, *Bearing Point*
- Jennifer Puyenbroek, Sylvia Shuler, Linda Johnson, David Merritt, Jeff Aycock, Anthony Morris, Jay Dalke, Wesley Stephens, Tommie Curtis, *SAIC*
- Leon Thorburn, Jean Thompson, *Northrop Grumman*
- Peter Hicks, *SRA International*

CS121852