Health Care and Informatics Review Online, 2008, 12(3), pp 37-44, Published online at www.hinz.org.nz

ISSN 1174-3379

Metadata Repositories in Health Care

Discussion Paper

Dr Karolyn Kerr
karolynkerr@hotmail.com

Abstract

Access to detailed information about the data held in the Enterprise Data Warehouse will assist in the overall management of data quality, including increasing the usability of the data. To make effective improvements to data quality the organisation needs to provide contextual information and clear standards on data definitions through a metadata repository that is adequately maintained and available to all stakeholders. Problems can arise where users and practitioners are not informed of the context of data collection and assumptions on the quality of the data can be made without access to sufficient data quality information. Most health care organisations now have complex data sets that require the support of a metadata repository. One approach is to utilise Total Data Quality Management to strategically manage data quality. By defining master data that is essential to the organisation, the DHB can then prioritise data quality improvements.

1. Introduction

This paper details current theory on metadata repositories and master data management. This theory is then applied to the local health care management environment, to highlight the need for more active metadata and master data management to support total data quality management (TDQM).

The benefits of a metadata repository include increasing the longevity of the usefulness of the data. Data users are provided with metadata that enable decision making on data in the data warehouse. The policies and procedures under which these data were collected may have changed, but with sufficient data quality information the user is able to utilise data from historical collections. Further, while the data may be accurate, if users do not understand the meaning of the data or the context of their collection, their interpretation of the data may be inaccurate (Olson, 2003). As many data users become increasingly removed from any personal experience with data, the importance of contextual information on their collection increases (Kerr, 2005).

A review by the author of the current literature available on metadata repositories did not elicit articles related to metadata repositories for health care management. References are limited to national repositories for the collection and storage of reporting data and for the management of clinical data and often related to clinical trials. However, the literature does provide general guidance that can be applied to health care management.

2. What is Metadata?

Meta is a prefix that in most information technology usages means "an underlying definition or description". Therefore, metadata is a definition or description of data. For example, a library catalogue is metadata describing publications in a library; a price list is metadata describing products for sale (Smith & Watmough, 2002). Information to be included in the metadata repository should include the quality of the information products extracted from the data warehouse.

Back room metadata is process related, and it guides the extraction, cleaning, and loading processes and is of use to database administrators and business users. The front room metadata is more descriptive, and it helps query tools and report writers function smoothly and are of use to end users. Front room and back room metadata can overlap (Kimball, 1998).

Without correct metadata, it is difficult to use the base data in an appropriate manner. Without easily accessible metadata, the challenge of using the base data becomes more significant with added effort to seek out the required metadata that is accurate, up-to-date, and complete.

2.1. Metadata Repository Models

Using the perspective of TDQM, Figure 1(Shankaranarayanan & Wang, 2007) provides an example of a logical model which can be useful as a subset of the metadata repository information. The model shows information products broken down into the elements included in each product and the processes required for input and output of data. Information products are associated with multiple product manufacturing stages. A data element is associated with one or more of these stages and these elements are assessed for quality measurements in the source data system. The metadata also captures the composition of each data element, in terms of other data elements that combine to make up the element. Time tags can track the elapsed time between data capture and the time the data becomes accessible in the data warehouse. Contextual weights can be applied by the data user for each specific data use requirement, meaning the user decides what the most important quality dimension in this instance is.

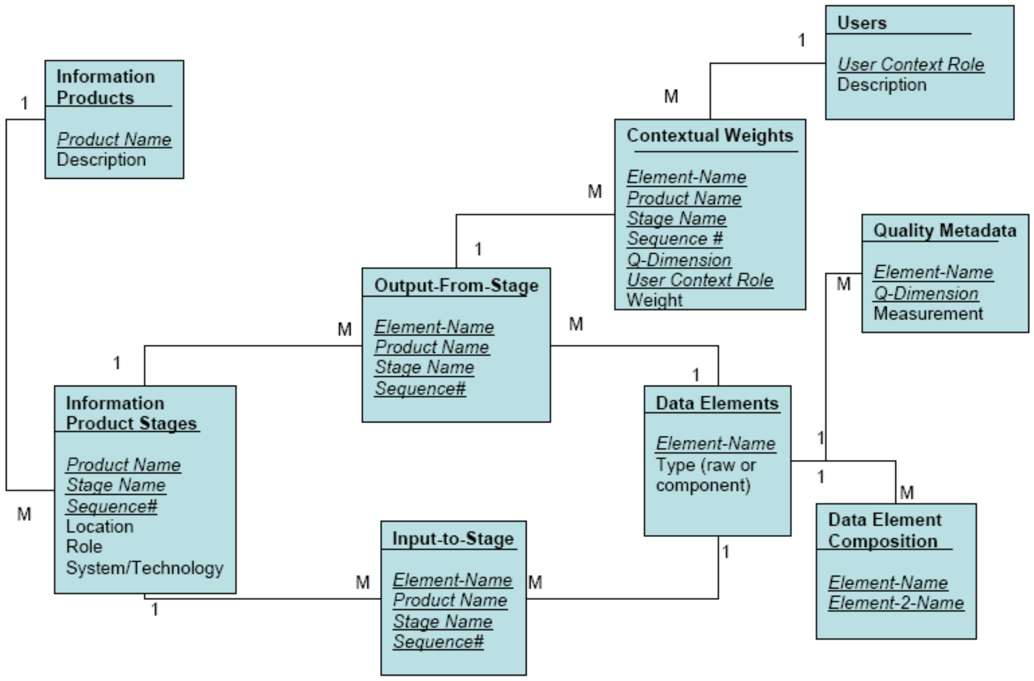

Figure 1: Logical Model of the Metadata Repository (illustrative sample subset is shown)

(Shankaranarayanan & Wang, 2007)

Figure 2 below shows a detailed ‘gold standard’ complete metadata model that includes all possible metadata from the organisation in one repository (Jennings, 2004). The data management sections of the model are represented by the enterprise systems component area. The purpose of the enterprise systems component is to define all of the data structures in the enterprise environment and describe the interrelationship of the associated metadata. This component is the base for all other model component areas. It is used to control and manage both data and metadata throughout the model. The enterprise systems component provides details into data, relationships, and the data movement and integration processes. The component is the reference source for system semantics and data structures in the model.

The management of an organisation's IT assets is represented by the IT portfolio management component area of the model. These assets include areas such as service management, software management, hardware and network management, project portfolio management and data quality management. The IT portfolio management component area supports IT management as a business function of an organisation. The objects of the IT portfolio management component represent real business items of an organisation but are integrated with the management of standard categories of metadata (e.g., databases, tables, flat files, etc.). The IT portfolio objects are planned and managed together with these more recognisable metadata objects.

The next component of the model represents the metadata associated with XML due to its increasing pervasiveness throughout organisations. Today, most applications in an organisation interact with XML at least at some level. These interaction levels include areas such as XML schemas (XSD), DTD, XML transformations (XSLT), business transactions and classification schemes. XML documents are used in an organisation for processing such as data transport between applications and Web services. The metadata used in conjunction with XML must be defined and used efficiently in order to avoid redundancy in an organisation.

The relationship of the business to information is represented through the business rules, business metadata and data stewards component area. The purpose of this component is to support the business user view of the metadata. The metadata represented in this component imitates the way the user interprets and uses this information. This component provides a standard method for documenting, maintaining and approving the rules, definitions and people that run an organisation. The data stewards of the organisation are depicted because they characterise the metadata needed for the business to make decisions. This business metadata allows users to organise data according to the way the business sees the information. The business relationships of data can be understood regardless of the physical structures or technology used.

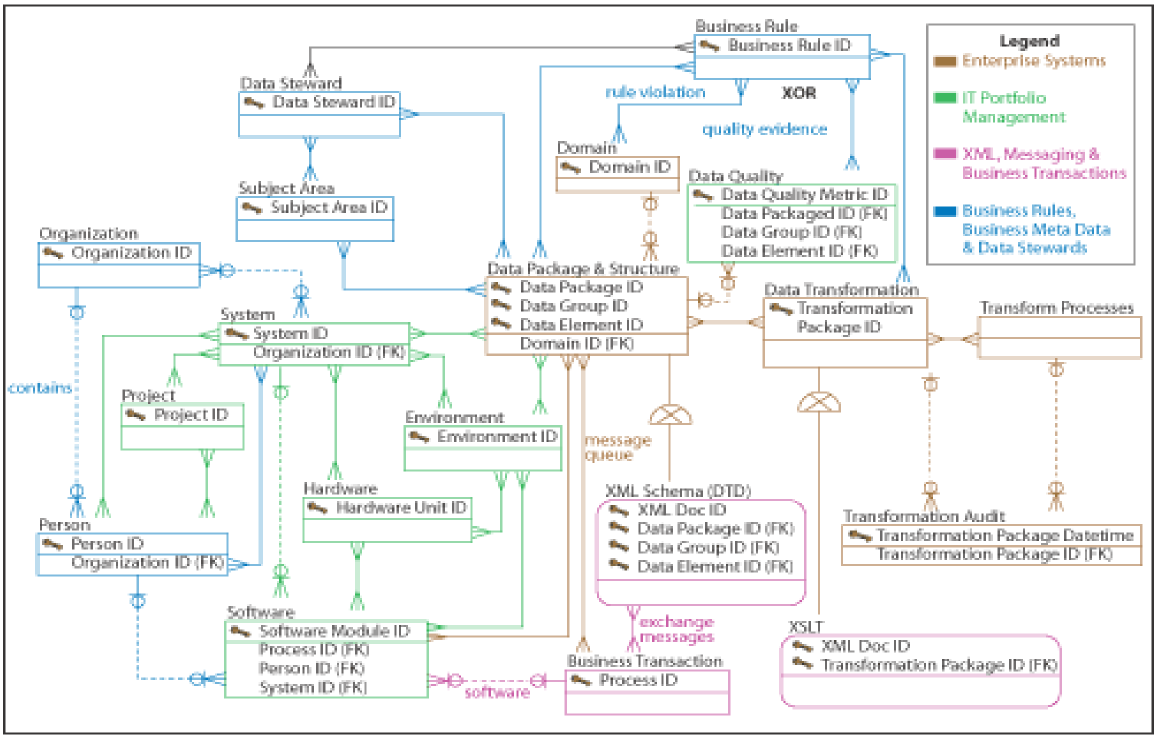

Figure 2: The Complete Meta Data Model (Jennings, 2004)

3. Master Data Management

Master data is the data that are critical sets of data in the organisation and usually consists of people (patients and staff), things (financial management, assets), places, reference data (such as the NHI and HPI), and does not include all of the organisations data. The master data repository becomes the standard source for all other systems, applications, and environments. Master data management centralises the capture, storage, and synchronisation of key business entities, and integrates the metadata of key business entities.

Master data management is defined as

> *A set of disciplines, applications and technologies for harmonising and managing the system of record and entry for the data and metadata associated with the key business entities of an organisation. (Powell Media LLC, 2006)*

The benefits of managing data in such a way include:

- Reduced master data redundancy
- More consistent master data
- Significantly more efficient business through a complete view
- Improved efficiency in change management processes
- Improved return on IT investment

Master data management is a multi-year evolving strategic initiative that includes the development of:

- A master data store and master metadata store – a repository used for storing and maintaining master data and its corresponding metadata
- Master data integration services such as enterprise application integration (EAI), extract, transform, and load (ETL) and enterprise information integration (EII)
- Master data management processes

What is important is that wherever possible, there is a common set of master data governance procedures and business models, and business rules to maximise master data quality. Clear policies and procedures around changes to master data are critical to ensure consistent data management.

Master data management development projects should be developed as an organisation wide strategic initiative with strong executive support. The challenges that are likely to be encountered in such an endeavour are similar to any other strategic data quality initiative, and include:

- Acceptance of ongoing roles and responsibilities
- Organisation-wide governance to ensure discipline
- Significant investment is required over time
- Significant human resources may be required
- Difficult to measure return on investment

The main capabilities required by an organisation wide master data management system can be broken down into five main areas, and are similar to that required by all data:

- Application design
- Metadata management
- Data management
- Integration services
- Data governance

Governance is required, and may be a role that is distinct from governance of data elements collected and used in individual departments. Master data is managed from an organisation wide perspective and requires governance at executive level.

4. Process for the Development of a Metadata Repository

The metadata repository should be developed to support total quality management across the DHB. Ideally, the repository data would reside in a single repository for consistent management of changes and updates.

Master data should be distinct, clearly defined by the business users and managed as a core dataset. Data that are used organisation wide and supported by a HISO standard could be included in this subset. The difference between managing master data from other data is that the definitions and coding used is ‘locked down’ and unable to be edited by business users, once there is agreement on a single definition. Ideally, business processes used to collect this data are the same and defined in the master data repository, such as a single form for data collection. Data obtained from outside the organisation, such as from PHOs, are difficult to manage in such a way.

Figure 3 below shows the relationship between master data, metadata, and the enterprise data warehouse (Shin, 2006). There is an overlap between master data management and metadata management as many of the management processes will be the same.

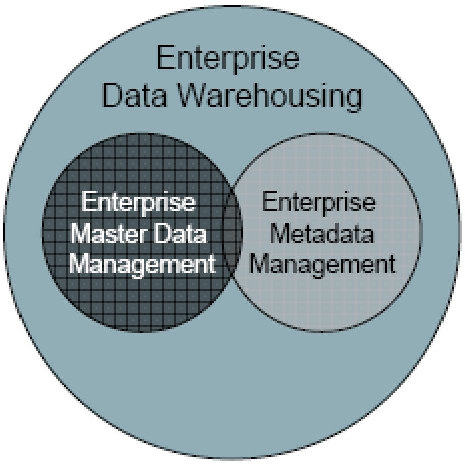
Enterprise Data Warehousing
Enterprise Master Data Management
Enterprise Metadata Management

Figure 3: Enterprise data management principles (Shin, 2006)

The metadata repository project should take a long-term strategic and iterative approach to development and an initial pilot may be appropriate. Architecture should take into consideration the evolving nature of the data in the organisation and the increasing amount of data elements and their metadata that will be added over time.

A core data set (master data) should be defined first, alongside department specific data that are essential business data. Alongside, business processes to administrate and maintain the repository are required. These processes include day to day operations and governance (structure, policies and procedures, roles and responsibilities) and stewardship for such areas as

- Change management
- Security management
- Operational support
- Performance management
- Stewardship workflow

4.1. What are the Requirements of the DHB Metadata Repository?

Metadata consumers require readily available access to different types of metadata, ranging from report definitions and their underlying columns, to various components of analysis associated with the data. The requirements for types of metadata can include:

- metadata for tables, views and files, and their supporting columns and indexes
- metadata for reports
- metadata for source to target mappings
- a facility for data quality and control metrics to be acquired actively through quality feeds
- external reference material
- notations and issues associated with various objects
- classification of repository objects into topical areas and subsystems
- audit trail of metadata changes within the metadata repository
- versioning of descriptions, notations and issues within the repository

- security may be a requirement where all those who have access to the repository should not have access to all areas of the repository
- custom web user interface via the intranet providing a direct access user interface

A logical and physical model of the data warehouse is required as a first step in the development process. The following steps are then required:

- make an annotated list of all possible metadata – this will include source system, data staging, DBMS metadata, and front room metadata
- prioritise the above list
- assign responsibility
- decide what constitutes a consistent and working set of metadata, what is master data, and what data are more flexible
- decide whether to develop in-house or buy a custom-made tool to manage the metadata
- develop the architecture required to store the metadata, including backup and recovery considerations
- develop tools to provide access to all possible users
- develop and implement data quality and maintenance procedures (Kimball, 1998)

The information that should be included in the metadata repository to support data quality is listed in table 1 below. The information includes historical and current information about a collection to ensure policies in place at the time of data collection are documented.

Table 1. Contents of the Metadata Repository

Reason Purpose for of the Collection This defines the purpose of the collection: what it is to be used for, why it is to be used in that way, any legislative and/or contractual requirements for reporting.
Coverage of the Collection Description of the population groups targeted by the collection
Stakeholders of the Collection Identification of the stakeholders whose activities will be supported by the collection. Description of the key information needs and reporting requirements of the stakeholders and how that information will benefit the stakeholders' activities. Identification of stakeholders who will be the source of data for the collection, and the relationship of the data collection to their activities.
Composition of the Collection Description of the data elements that comprise the collection, with an associated data dictionary.
Access, Privacy & Security Requirements Description of how the stakeholders will access the collection, and how these access arrangements will protect the privacy of the people who provided the data in the first place. Description of security controls (including authenticating the person seeking access to the collection) and how the effectiveness of the controls will be monitored.
Key Process Flows Description (including process maps) of the process flows for collection, production and maintenance of data. This includes the timeliness required from this process as well as the process itself. Processes also include the extract, transform and load processes

Business Rules & Data definitions Definition of this defines the business rules applied to the preparation of data which also validate and transform data received.
Quality & Audit Requirements & Quality Assessment Definition of what is acceptable data quality within the collection (including all dimensions of data quality Definition of which business units are responsible for maintaining the quality levels at different stages in the data flow. Description of the mechanisms for assessing quality and reporting on it.
Governance This includes a description of the stewardship and custodianship arrangements. It also includes a description of the operational funding for the collection.
References to Standards Identification of the standards relevant to elements of the collection.
Future Requirements Request for additional functionality planned for the collection.
Collection Awareness/Communication Definition of the level of awareness and communication to be maintained across stakeholder groups.

Data that are specific to individual departments, or defined differently by departments, may also be included in the metadata repository. This data may initially make up the majority of the metadata held in the repository. Over time, as data elements are defined organisation wide, it can be moved into the master data repository.

A web-based interface to the metadata repository will allow all appropriate users access to the information. Data requests can also be made through this web interface. A search browser will also assist users to find information, particularly once the repository increased in size.

5. Conclusion

This paper outlined the current theory on metadata repositories and master data management to support improved data quality. Given the now complex nature of much of our healthcare data, a structured and prioritised approach to managing metadata is required. Master data management using a TDQM approach allows for the management of core business data in a consistent way across the organisation, including business processes, governance, and change management.

6. References

[1] Kimball, R. (1998). *The Data Warehouse Lifecycle Tookit*. New York: John Wiley & Sons.

[2] Olson, J. E. (2003). *Data Quality. The Accuracy Dimension*. San Francisco: Morgan Kaufmann Publishers.

[3] Powell Media LLC. (2006). *Master Data Management: Creating a single view of the business* (Research Report): Business Intelligence Network.

[4] Shankaranarayanan, G., & Wang, R. Y. (2007, November). *IPMAP: Current state and perspectives*. Paper presented at the 12th International Conference on Information Quality, November 2007, pp. 510-517., Boston.

[5] Shin, M. (2006). *Master data management: A framework and approach for implementation.* Paper presented at the 11th International Conference on Information Quality, Boston.

[6] Smith, D., & Watmough, W. (2002). A successful approach to a metadata repository.

[7] Kerr, K. (2005). *The Institutionalisation of Data Quality in the New Zealand Health Sector*. PhD Thesis. Auckland University.

[8] Jennings, M. 2004. *The Complete Metadata Repository*. DM Review. www.dmreview.com/issues/20040401/1000941. Accessed 12.6.08