A Data Quality Handbook for a Data Warehouse

Eva Gardyn, HSBC Asset Management, Melbourne, Australia.

Abstract

This paper provides a draft outline of a data quality handbook for a data warehouse. Data quality is based on the intended usage and application of the data and is demonstrated by a data warehouse that is used to provide overall company reporting and decision support. A data warehouse reuses data that was originally captured for another purpose. The same data held in both a data warehouse and in other systems can have different quality criteria, depending on how the data is intended to be used. This paper proposes five data quality dimensions to which properties of a data warehouse can be linked. This data quality handbook identifies the properties of a data warehouse that have a direct bearing on the quality of data in the data warehouse.

1. Introduction

As organisations require and capture more data, the quality of the data stored has become a major industry concern. Organisations are relying more on data in data warehouses to make critical decisions about their operations. Date warehouses are becoming very popular in industry, as most large organisations either have already implemented one or are in the process of doing so.

Recent surveys of several large corporations as documented by Mattison [Mattison96] reveal that business are planning on spending large sums on developing data warehouses. The sample survey included organisations across all industries, including manufacturing, financial, health, retail, utility, government, banking and telecommunications fields. Of those firms surveyed, 90% were involved in some stage of data warehousing systems development.

Some information about these data warehouse projects include the following:

- Of the organisations working on data warehouse projects, over 60% are planning on storing more than 20 gigabytes of data and over 10% plan to build systems that involve more than 300 gigabytes.
- Budgets for these projects range from $250,000 to several million dollars.
- The numbers of users to be supported range from tens to thousands.

The concept of a data warehouse arose to allow organisations to report on company-wide data that are captured in the systems that operate the business (the operational systems). A separate system to the operational systems was needed to provide reports on the operational data because of the difficulties in extracting the required data from the operational systems and because of the risks involved if reporting was to interfer with systems that process critical business operations. A data warehouse provides reporting and decision support based on operational data (and other data) without interfering with the business operations. The purpose of a data warehouse is to make data available to the users via a single interface where that data would otherwise be difficult for the users to access and use.

A data warehouse reuses the same data that was originally captured for another purpose and the same data held in both a data warehouse and in other databases, can have different quality criteria, depending on the use of the data.

1.1 Quality Model Framework

Most of the literature on quality has focused on the quality processes. Dromey [Dromey96] has described a way to measure software *product* quality and identified a fundamental axiom of software product quality as "*tangible internal characteristics or properties of a product determine the external quality attributes that it exhibits*".

This fundamental axiom can also be applied to data quality in a data warehouse and is paraphrased, as follows: "*tangible internal characteristics or properties of a data warehouse determine the quality of data provided*". (I have interpreted Dromey's term "tangible" to mean observable and measurable).

Dromey in [Dromey96] defined a Quality Model Framework that he used to link tangible software product properties to high-level quality attributes. This framework is used here to link data warehouse properties to data quality dimensions.

The framework has the following three principal components:

- a set of data quality dimensions
- a conceptual model for a data warehouse which identifies the properties of a data warehouse that influence quality,
- a means of linking observable data warehouse properties to the data quality dimensions.

Figure 1 depicts my application of Dromey's Quality Model Framework for data quality in a data warehouse.

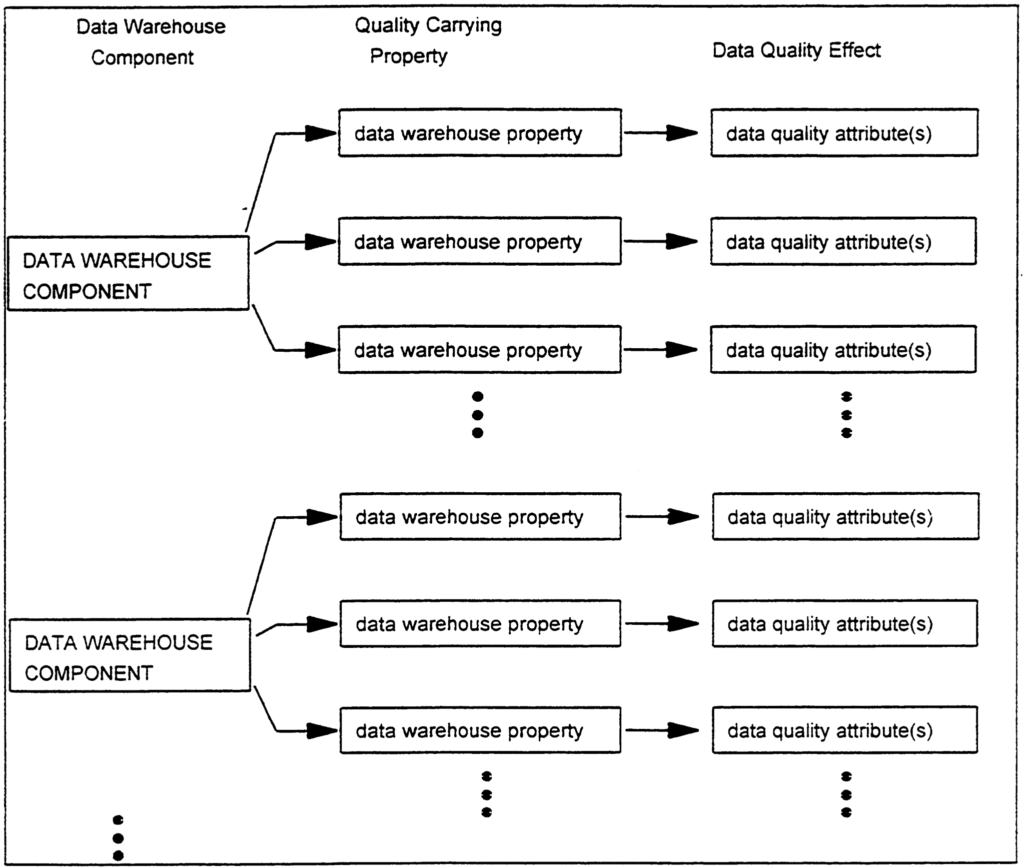

Figure 1. Quality Model Framework for data quality in a data warehouse.

The Quality Model Framework showing actual data warehouse components, properties and links to actual data quality dimensions is developed in Section 4 and shown in Figure 5.

1.2 Contents of Paper

A data warehouse is described in Section 2. Although there is no "official" definition of a data warehouse, I have provided a definition and description of the main components of a data warehouse, based on various definitions in the main literature about data warehousing and my industry experience. Section 2 includes a description of the purpose of the data in a data warehouse as well as descriptions of the components of a data warehouse. Section 3 proposes a set of Data Quality dimensions to which data warehouse properties will be linked. A set of observable processes and features (the quality properties), that influence data quality for each of the data warehouse components, is defined in Section 4.

2. A Data Warehouse

There is no "official" definition of a data warehouse despite the abundance of information about data warehousing. The following description has been obtained from the main literature about data warehousing [Inmon92], [Mihulka95], [Devlin&Murphy88], [Percy95], [Poe96], [Mattison96], [Kimball96], [Brackett96] and my experience in the practical use of data warehouses in business today.

A data warehouse is a system consisting of processes and databases used to provide end users access to integrated company-wide data for decision support. Processes extract data from various source systems, databases and files, and transform the data into an integrated, consistent format. The integrated data may be summarised and manipulated and are stored in a read-only database. The users access the data via various front-end tools and decision support systems (DSS). Section 2 describes a data warehouse and shows the architecture of a data warehouse, which consists of components for data acquisition, storage and data access. An Information Catalogue is an important component of a data warehouse and it provides the meta data to support data acquisition, storage and access.

2.1 The Purpose of Data in a Data Warehouse

Data quality is a direct extension of quality management [Deming86] where quality was defined as *fitness for purpose*. Data quality depends on the use of data [Wang,Reddy,Kon93], [Strong,Lee,Wang97] and the same data may have different quality criteria, depending on the use of the data. Most of the data in the data warehouse, used for management reporting, are derived from existing data bases (the operational systems) where the data in the operational systems was originally captured to process critical business activities.

Data provided by a data warehouse has been described in various ways in the data warehouse literature [Brackett96], [Inmon92], [Mattison96] and others. These descriptions have included the terms corporate, strategic, organisational, institutional, analytical, transformational, directional, planning, enterprise and informational.

A data warehouse uses the same data that is also stored in the operational databases. The operational databases originally captured the same data for very different purposes and provide this data to the data warehouse for decision support and management reporting. Brackett [Brackett96] distinguishes between the different usage of data in the operational systems and data warehouse as *operational* versus *evaluational* usage. The data in operational systems are used to *operate* the business, whereas the data in data warehouses are used to *evaluate* the business. The purpose of the data in a data warehouse is to provide decision support and various management reporting on an organisation's operations. This includes evaluating the current status, trends and patterns about an organisation's business and making projections to analyse future alternatives. The users of a data warehouse use the data to manage large sections of an organisation, or even the entire organisation.

Data within a data warehouse are not (usually) captured specifically for a data warehouse. The data is input to an operational system for a specific application (with its own data quality criteria) and then provided to a data warehouse for a different application, namely management reporting, that has different data quality criteria.

The purpose of data in a data warehouse is to provide management with company-wide information for decision support and management reporting. It is used for business planning, marketing and similar functions, where the data are used to provide information such as overall trends of company performance and customer segment profitability. A data warehouse is designed to satisfy queries about the entire business and enable cross-department and cross-product reporting. In contrast, operational systems are designed to process transactions and therefore hold data specific to that purpose. The quality of data in the operational systems may be adequate for their original purpose but can cause problems when the same data is used to source a data warehouse.

An Example: Same Data, Different Purpose

A current account system is a typical operational system in a bank, which processes current account transactions and produces a statement of the accounts' transactions and balances at periodic intervals. This system stores certain data for this purpose, including the customer name and address. The customer name and address are required for the delivery of the bank statement to the customer and are considered of adequate quality if the data enables the statement to be delivered to the correct customer. In the current account system, the purpose of name and address data is *to identify to Australia Post the correct delivery point for the bank statement.*

The name and address data may typically be stored as free-form unstructured text which is all that is required for the delivery of the statement to the customer.

An example of the name and address fields that can be held in the current account system is:

NAME-AND-ADDRESS-LINE1
NAME-AND-ADDRESS-LINE2
NAME-AND-ADDRESS-LINE3
NAME-AND-ADDRESS-LINE4
NAME-AND-ADDRESS-LINE5
NAME-AND-ADDRESS-LINE6.

(Note. These fields show that the data is stored in six unstructured strings of text.)

The same name and address data is input to a data warehouse and is required to serve a different purpose. It is used to obtain a consolidated view of customer across different lines of business to provide answers to queries such as report on all customers where total income over the past quarter was greater than $50,000. The purpose of the address is *to provide for trend analysis across geographic lines and to report on relative profitability across different states.* The data in a data warehouse needs to be represented in a consolidated form and business entities need to be represented in the same way. Data cannot be buried in free form text if it needs to be analysed.

An example of the data structure that would be required in a data warehouse for name and address is:

SURNAME / ORGANISATION NAME
SUBURB
STATE

In this example, the address data in the current account system is good, if the address data can be used to successfully deliver bank statements to customers and the data structure is irrelevant. However, in a data warehouse, the data that was perfect in the Current Account

System is now considered of low quality because it is badly structured. In a data warehouse data structure is important and contributes to the data dimension of *accessibility*. (See Section 4). Inappropriate data structure contributes to poor data quality.

2.2 Distinctive Characteristics of a Data Warehouse

The fundamental purpose of a data warehouse is to provide integrated data to end users for decision support and management reporting. The data that is provided usually resides in various databases and systems throughout the organisation but is not easily accessible in an integrated way. A data warehouse does not provide any functionality other than extracting the data (that has already been captured), integrating the data and presenting the data to end users. There is little processing involved in a data warehouse application. The complexity is in integrating disparate data from multiple sources to provide various facets of company-wide information.

The applications and users of a data warehouse require access to company-wide data for management reporting. The characteristics of such applications differ significantly from the characteristics of operational and safety critical systems in a number of ways.

These include:

- no need to have real-time currency of data
- precision of the data will not be critical
- access is read only, no updates are made
- ability to view and aggregate the same data in different ways
- "snapshots" and historical data spanning a number of years, are required.

2.3 Components of a Data warehouse

A data warehouse is a system consisting of processes and databases used to provide end users access to integrated company-wide data for decision support. Processes extract data from various source systems, databases and files, and transform the data into an integrated, consistent format. The integrated data may be summarised and manipulated and are stored in a read-only database. The users access the data via various front-end tools decision support systems (DSS). Figure 2 shows the architecture of a data warehouse, which consists of components for data acquisition, storage and data access. An Information Catalogue is an important component of a data warehouse and it provides the meta data to support the data acquisition, storage and access components.

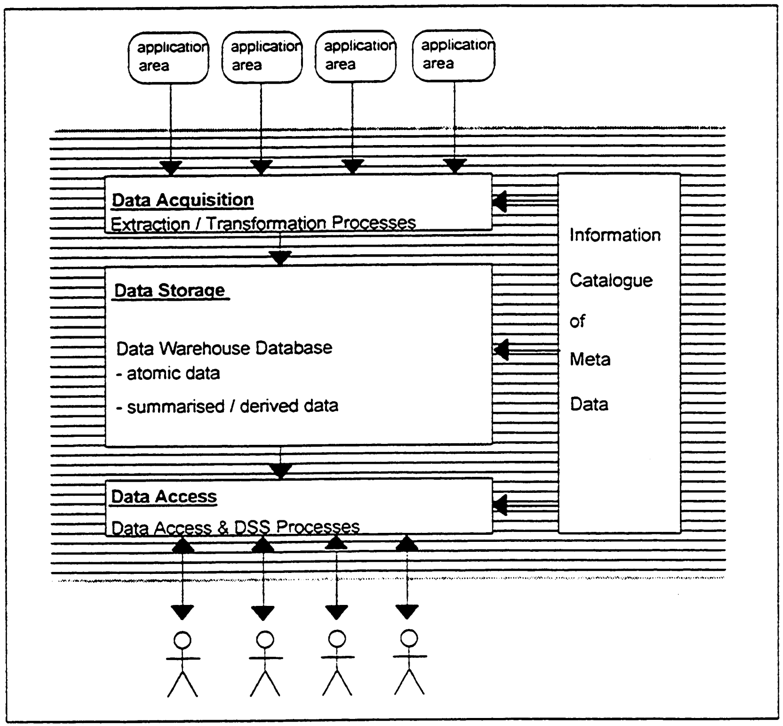

Figure 2. Architecture of a Data Warehouse

The data warehouse components are as follows:

- data acquisition component
- data storage component
- data access component
- information catalogue of metadata component to support data acquisition, storage and access.

2.3.1 Data Acquisition

The Data Acquisition component includes all the programs, data staging areas and operational system interfaces that extract data from mainly disparate operational systems and preparing the data for the data warehouse, prior to loading the integrated data into the data warehouse database.

The source systems are usually on-line transaction processing (OLTP) systems that provide operational support for running the day-to-day business of an organisation. For example, in a bank an OLTP system that will source a data warehouse will be the system that maintains and updates bank accounts. The data items that are necessary for decision support and company-wide reporting are extracted from the source systems. Other data sources may be external to the organisation such as companies that specialise in providing financial market data (e.g. Reuters or Telerate). The sources for a data warehouse may be on different databases, different platforms and in a variety of data types and formats.

There is usually a need to transform the data in some way. For example, the Customer Identifier in the Cheque Account system may be different from the Customer Identifier used for Credit Card application. The processes remove redundant, coded data and may also "cleanse" the data during the transformation, by correcting any errors encountered, based on data validation rules established for the data warehouse.

[Brackett 96] identified the processes that should be undertaken when formally transforming disparate data into an integrated data resource. These steps are also appropriate for the Data Acquisition component of a data warehouse environment as the Data Acquisition component extracts data from disparate sources (the operational systems) into an integrated data source (the data warehouse database). The processes included in the Extraction/Transformation component are:

Identify Best Source Data.

Where multiple sources of the same data exist across operational systems, the "official" source of data must be identified. The official data source is the best source of data and may differ from the primary source. (The primary source is the point at which data was first captured).

Extract Source Data

Source data are extracted from operational systems prior to further data transformation. The extraction includes:

initial data extraction - to build the data warehouse database, and
ongoing data extraction - to provide snapshots of operational data on a regular basis.

An interim, "staging" database may be necessary where data is being extracted from one physical environment of the operational system to go into a different physical environment of the data warehouse database.

Translate Data to Consistent Form

Disparate data from operational systems are transformed into an integrated, consistent form for the data warehouse. This will include standardising codes, field lengths, names, and so on.

Recast Data

Recast data for historical continuity where different data structures exist over time. This is required when the operational data structures such as primary key, format, etc. change over time. e.g. may be as a result of changes in a corporations organisation structure, changes in product identifiers, and so on.

Restructure Data

Alters the structure of the data from the disparate operational data to the integrated data warehouse data. Includes both logical and physical restructuring. Logical restructuring establishes the appropriate "normalised" form for the data. Physical restructuring may be required to denormalise data to be operationally efficient in the data warehouse database environment.

Summarise Data

Summarises the operational data to the desired level for the data warehouse. Need to identify the most detailed level required for drill-down and other data access requirements. The desired level of summarisation depends on volumes of data, type of analysis desired and the time constraints to perform the analysis.

Load Data

Loading the data includes any further data editing which provides checks on data entering the data warehouse, to ensure conformance with business rules and avoid low-quality data entering the data warehouse. Procedures need to be established to ensure data that failed the edit check is passed back to the operational system for correction and reload.

2.3.2 Data Storage

The database of a data warehouse holds the data that will be used by end users and other applications. Unlike most operational databases in an organisation, for the end users, a data warehouse database is read-only - "you can't change the past". The updates to the database are only from the source systems and will typically occur at fixed intervals (e.g. at the end of day). Usually, the database will hold a range of historical data so that users can monitor trends and information patterns over time. The database will be optimised for information access and the design will usually be not as fully normalised as the designs of operational databases.

A major design issue is establishing the appropriate level of granularity to efficiently process the vast majority of requests but also provide answers for any ad-hoc requests on the company data. This requires a trade-off between volumes of data stored against the level of detail that can be provided for a query. A data warehouse database will usually have at least two levels of granularity.

Inmon [Inmon 92], Brackett [Brackett96], and Kimball [Kimball96] have identified characteristics of data warehouse databases and shown how these differ from characteristics of operational databases. These differences in database characteristics are shown in figure 3, below:

Data Warehouse Database	**Operational Database**
Integrated	Non Integrated
Subject Oriented	Process Driven
Non-Volatile	Dynamic
Time-Variant	Single time
Different levels of granularity	Primitive and detailed granularity
Denormalised	Highly normalised

Figure 3. Differences between operational and data warehouse databases

Integrated

Data is stored in the data warehouse database in a singular, company-wide fashion, even when the underlying operational systems store the data differently. The integration of the database includes consistent naming conventions, consistent measurement of variables, consistent encoding structures, consistent physical attributes of data, etc.

Subject Oriented.

The data warehouse database is organised by data subjects based on the identification and definition of objects and events that are of interest to an organisation. This data-driven approach is in contrast to the process orientation of many operational systems.

Non-Volatile

Operational systems are considered dynamic and are updated (inserts, deletes, changes) on a record-by-record basis on a continuous and (usually) random basis. In the data warehouse, data is loaded at scheduled periods and no further updates are made. The data is accessed as read-only.

Time-Variant

In the data warehouse, data is typically included for a long time horizon (5-10 years) and consists of a long series of snapshots each of which refers to some moment in time. The data element is (usually) time-stamped to indicate the time period it refers to. In contrast, the operational systems (usually) have data that refer to the moment of access and do not hold data that refers to other periods of time.

Different Levels of Granularity

A data warehouse will usually contain data at different levels of granularity, including primitive, detailed, derived and summary. The appropriate levels of granularity are determined to ensure efficient processing of the majority of requests as well as provide answers to ad-hoc requests. Data in operational databases are held at a primitive and detailed level of granularity.

Denormalised

Data warehouse data is not fully normalised like operational data as processing efficiency requires minimum navigation of data subjects. Normalisation is necessary in operational databases to avoid update anomalies but is not necessary for a read-only data warehouse database. Levels of denormalisation includes holding derived data, allowing repeating groups (first normal form) and interattribute dependencies (third normal form) to be left unnormalised.

2.3.3 Information Catalogue

The information catalogue provides details about the data within a data warehouse database. These details include:

- definition of the data (semantics)
- structure of the data and a record of the changes of the data structure
- mapping between one or more legacy data and a data warehouse datum
- algorithms used to convert, enhance and transform data from the legacy systems to a data warehouse database

- quality profile of the data.

[Brackett96] calls the Information Catalogue for a data warehouse *metadata.* He describes the need for the Information Catalogue to provide an awareness and understanding of the data so that it can be fully utilized to support business needs.

[Brackett96] has suggested that the Information Catalogue include Data dictionary, Data Structure, Data Integrity rules, Data Thesaurus and Data Translation Schemes. I have also included a Data Quality Profile which is similar to the tagging of data with relevant indicators of data quality proposed by Wang, Reddy and Kon in [Wang,Reddy,Kon93].

Data Dictionary

Contains formal data names and comprehensive data definitions for:

- data sites,
- data subjects,
- data characteristics
- data characteristic variations
- data codes
- data versions

Data Structure

Data structure contains the structure of the logical data of the data warehouse data model. Data Structure includes:

- data subject
- data characteristic
- data characteristic variation
- primary key
- primary key characteristic
- foreign key

Data Integrity Rules

The data integrity rules document rules for maintaining the data integrity of the data in the data warehouse. These rules are defined for data subjects and data characteristics. These rules are either a rule or a set of allowable values. Can also have conditional data value integrity which shows allowable combinations of data values. Also includes a set of characteristic options which specify mandatory and optional requirements.

Data Thesaurus

Contains a set of data name synonyms and provides a reference between similar names or business terms & common data names.

Data Translation Schemes

Provides a translation between common data characteristic variations that represent measurements or format variations.

2.3.4 Data Access

The Data Access component includes the processes and tools that access summaries of company-wide data and include "drill-down" analysis allowing navigation from summary to base data. The processes and tools support trend analysis through the use of historical data. In contrast to most operational systems which process one record at a time, the typical usage in the data warehouse involves large summaries of records.

Multidimensional analysis tools support the analytic requirements of the data access and DSS component, by allowing on-the-fly calculations and summaries of data in the data warehouse. These tools provide a set of computational and data-navigation capabilities that derive answers to queries based on data stored in the data warehouse. An example of a typical query such as "How have advertising expenditures affected sales?" will be derived from data stored in the data warehouse and presented as a report that can be navigated by the user. The user can drill up, down and across the report and change report parameters as required.

Mattison [Mattison 96] and Kimball [Kimball96] have identified the features that should be part of the Data Access component. These are not discussed here as Data Access is not included in the analysis of data quality contributors.

3. Data Quality Dimensions

Quality has been defined as "the totality of features and characteristics of a product or service that bears on its ability to satisfy given needs" [IEEE83]. Fundamentally, quality depends on the needs to be satisfied. This also applies to data quality where the quality of the data depends on its use. Data quality for a safety critical system will be significantly different from data quality for a marketing system.

Many data quality dimensions have been defined in the Data Quality literature [Agmon&Ahitov87], [Redman92], [Fox,Levitin,Redman94], [Wand&Wang96], [Orman,Storey,Wang96], [Cykana,Paul,Stern96], [Wang&Strong96], [Strong,Lee,Wang97]. For illustration of the concepts in this paper, I have taken dimensions that most researchers consider are important and that are relevant to data in a data warehouse. This could be extended to a more expansive set of data quality dimensions.

The data quality dimensions are described below.

3.1 Correctness

Correctness has been identified and cited [Wang,Storey,Firth95] as the most important dimension of data quality. The Collins dictionary defines *correct* as "free of error, conforms to a standard". Other definitions for *correctness* have included *accuracy* and *precision*.

[Wand&Wang96] identified correctness as an *intrinsic* dimension of data quality, where an absence of correctness is when "a state in the real world system is mapped to a wrong state in the information system." I believe that their definition is not appropriate as the *real world* system is only that which is perceived by the users and thus the "real world" on which an information system is modelled. is an abstraction based on perceptions / needs of the users. I differ from

[Wand&Wang96] by stating that all data quality dimensions depend on the *use* of the data as specified by the users.

Correctness (and incompleteness and inconsistency) can only be defined with respect to a specification. Therefore, I define correctness as:

The extent to which the data matches another set of data which acts as a specification or a reference set and the extent to which the data conforms to the business rules as specified by the users.

This set of data may be some aspect of the "real world", such as the addresses of clients, to be obtained from visiting the clients on a particular date. Or the data set may be in a file or another computer system.

3.2 Completeness

Technically, completeness (and consistency) are a part of correctness. Incompleteness problems in a given data set point to possible correctness problems. I have treated correctness, completeness and consistency separately as they have often been cited separately as important data quality dimensions.

Data is considered complete if all necessary values are included so that user information requirements can be met. "A set of data is complete with respect to a given purpose if the set contains all the relevant data and all mandatory attributes should be non-null (and within the prespecified range)." [Redman92].

However, [Wand&Wang96] define completeness as an intrinsic data quality dimension where "each lawful state in the real world system is mapped to a lawful state in the information system ..and is incomplete if there are lawful states in the real world system that cannot be represented by the information system." As also stated by Kent [Kent78], given the complexity of the real-world, I don't believe that it is possible (or useful) to identify *all* lawful states of the real world.

I use the following definition for completeness in a data warehouse:

All data is available to satisfy the user requirements that a data warehouse is intended to satisfy.

A data warehouse may serve different users with different requirements who may therefore correctly ascribe different completeness measures to it. This also applies to other data quality dimensions.

3.3 Consistency

Consistency has been defined as "entity types and attributes should have the same basic structure whenever possible". [Redman92]. Where there is only a single representation of data, it will be consistent. Problems with inconsistency arise when we have more than one representation of the same data. In a data warehouse, data is represented as captured in the source system, and may also be summarised or manipulated in some way. The value of the data in these two instances must be the same for the data to be consistent.

[Wand&Wang96] define inconsistency where there is more than one state of the information system matching a state of the real world system and that users can infer the real-world system from the representation in the information system. [Wand&Wang96] therefore do not consider *inconsistency* as a deficiency, based on their assumptions for data quality and do not include consistency as an intrinsic data quality dimension.

However, [Wand&Wang96] have distinguished between consistency and lack of ambiguity. They define ambiguity as "if several states in the real world system are mapped into the same state in the information system, there is insufficient information to infer which state in the real-world system is represented" and consider *lack of ambiguity* as an intrinsic data quality dimension. In a data warehouse, we limit the states (or data values) to those that are necessary to meet user information requirements. Ambiguity as defined by [Wand&Wang96] is not relevant for data quality in a data warehouse. Two different states of the real world-system that are mapped to the same state (data value) in a data warehouse will either be incorrect in a data warehouse or acceptable, depending on the use of the data. For example, if we record the country of the customer as "Australia" or "Overseas" and a customer from England and a customer from China both have their different countries recorded as "overseas". If this results in user information requirements not being met as they need to differentiate between different overseas countries, then the data is incorrect. If there is no need for the users to differentiate between different overseas countries, then the data is correct.

The definition of consistency in a data warehouse is:

There is a single representation of the same data, or if more than one representation, copies are controlled and have same format and content. The format and content conform to the business rules as defined by the users.

3.4 Currency

Currency has often been cited as a desirable data quality dimension [Wang,Storey,Firth95] and has been used to define how up-to-date a datum is [Redman92]. In the data warehouse we are likely to hold historical and time-series data. Also, "current" data in a data warehouse is not (normally) real-time. Current data may (validly) mean data that is one day old. or even one month old, depending on the user requirements. In a data warehouse. the desired currency of the data is that which is specified by the user, and may vary from data warehouse to data warehouse.

Wand and Wang [Wand&Wang96] have not included *currency* as an intrinsic data quality dimension but have recognised it as an external data quality dimension.

I define currency in a data warehouse as:

Data is up to date according to the user-specified timing. This may be as of close of business previous day, as at end of last week, etc. This definition for currency also applies to time series data and historic data.

3.5 Accessibility

Users know all data that is available in the warehouse and are able to access this data. This allows users to have a complete and consistent understanding of the data available to them. The absence of full definitions may be a major cause of poor data quality because users may have different interpretations of the same data. Inaccessible data may be due to inappropriate data structures. For example, a post code embedded in a free-form string is not easily accessible.

The definition of accessibility in a data warehouse:

All data is completely defined (the meta data) in terms meaningful to the users and these definitions together with the data are available to the users. Meta data includes source, meaning and derivation and data have appropriate data structures.

4. Data Warehouse Properties That Affect Data Quality

As described in the introduction, a data warehouse differs from other conventional data bases in a number of ways. One significant difference that affects data quality concerns the capture of data from the external world (its primary source). A data warehouse cannot control the capture of data from the external world into the organisation, when it is input to the operational databases. A data warehouse only has control from the point of the extraction process of the data from the operational systems into the data warehouse.

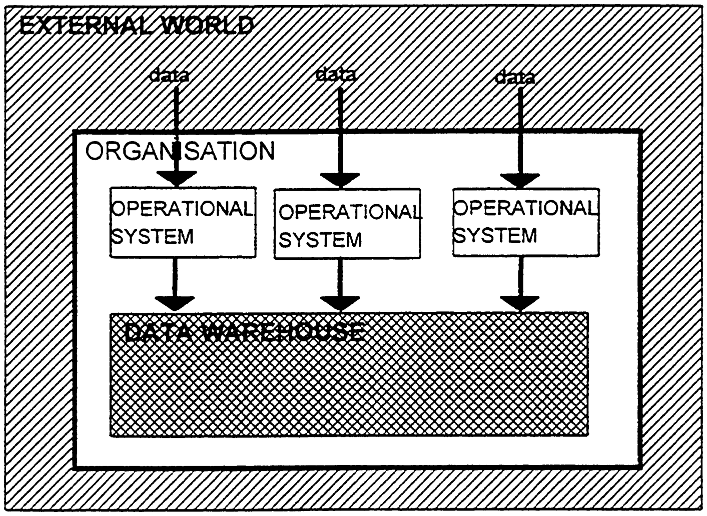

Figure 4. A Data Warehouse & its relationship to the external world.

However, despite incomplete control over the data capture process, the components of a data warehouse can be designed and implemented so that they affect the quality of the data held in a data warehouse.

The following sections describe a set of observable processes and features that can be designed and implemented in each major component of a data warehouse, that have an influence on the quality of the data. These processes and features are referred to as the *quality properties* of the data warehouse.

Quality properties are identified in the following sections for the Data Acquisition, Data Storage and Information Catalogue components of a data warehouse. The Data Access component, which includes the user interface design, has been excluded from this analysis as it is primarily concerned with presentation of data to the end user. Although the presentation of data to the end user and the overall design of the user interface can have a significant effect on the usability and therefore the quality of the data [Borenstein91], it is beyond the scope of this paper. Much has been written on the subject of good user interface design and its affect on the usability of the information system (see [Borenstein91], [Heckel91], [Norman88]), which can also be applied to a data warehouse.

For each of the data warehouse components, Data Acquisition, Data Storage and Information Catalogue, the links between the quality properties of the data warehouse component and the data quality dimensions are shown.

This section can be viewed as an initial draft for the outline of a data quality handbook for a data warehouse as it describes the processes and features that can be designed into and implemented in a data warehouse and that will have a positive influence on the quality of data in the data warehouse.

4.1 Quality Properties of The Data Acquisition Component

The following section provides an outline of the quality properties of the Data Acquisition component of a data warehouse. These quality properties are processes or features of the Data Acquisition component that can be designed and implemented. A detailed description of each quality property is beyond the scope of this paper, but the first quality property, the process to "Identify Best Source of Data" is described further to illustrate this idea.

4.1.1 Identify Best Source Data.

This is a *process* within the Data Acquisition component that is necessary when there is more than one physical representation in the operational systems for the same data and these different physical representations are not synchronised (not managed copies).

The best source of data is chosen because the data in that source may be more correct, current, secure, etc. than in the other sources. The details described here have been obtained from the processes established in a large Australian telecommunications company.

This process consists of identifying candidate sources and evaluating these sources against criteria specified. The best candidate is selected as the official data source. Criteria used in this process include:

- fitness

- content
- maintenance
- data structure
- accessibility
- future plans.

Evaluate Candidate Source for Fitness

This subprocess assesses how well the data in the candidate databases matches the required data in terms of meaning and completeness of coverage.

Assess Data meaning	Check contents of data to ensure it really is the required data.
Assess Completeness of Coverage	Identify if all pieces of the required data are in the data base

Evaluate Candidate Source for Content

Assess the perceived accuracy of the data values in the candidate sources. Check usability of the values in the primary keys and foreign keys for linkage to other pieces of data.

Accuracy of data values	Evaluate user perceptions of data accuracy. Check for any known data values problems or misinterpretation.
Validation Procedures	Check if validation procedures exist in the application. Check if there are cross checking transactions or reference tables / values. (e.g., subtotals, valid ranges, selection lists of valid values).
Usability of Values in the Primary Keys	Check whether referential integrity has been implemented. Check that primary keys are *unique* identifiers of the required data.
Linkages to other data subjects (applications)	Check whether data and identifiers are used by other applications.

Evaluate Candidate Source for Maintenance

Assess staff expertise and training of maintenance functions. Evaluate security levels and appropriate staff assigned to maintain data.

Staff training for data maintenance	Check complexity of the process for data maintenance (create, delete, amend data). What is the level of automation? Establish appropriate level of training and skills that is required by data entry staff.
Access to maintenance procedures.	How secure is the data maintenance procedure? Establish whether suitable people have appropriate access.
Number of transactions required to maintain data.	Determine number of business transactions and TP monitor transactions required.

Number of data maintenance transactions per day.	Obtain transaction volumes per day/hour/second.
Number of batch jobs creating data	Obtain batch volume processing, file transfers, data aggregation, etc.
Duration of create batch jobs	Obtain duration time, frequency of jobs, significant sequence of jobs, triggers.

Evaluate Candidate Source for Data Structures

Evaluate the flexibility of data structures for future development.

Primary Key Structure Consistency	Check that key identifiers of data are consistent across application.
Primary Key Formats	What is the format of the primary key: attributes & definitions.
Meaning in Primary Key	Are primary keys meaningless and thus independent of the data they are identifying? Data content in primary keys means limitation and dependency.
Follows Standards	Do the data structures follow corporate standards?

Evaluate Candidate Source for Accessibility

Assess the options available to access the data and costs associated with different options.

Hardware Platform	Obtain platform details, check limiting factors, size and volumes
Network	Types of network connections available to the application and access to data. Obtain data transfers, volumes, bottlenecks, gateways.
Transaction Monitor	What transaction monitors are used? (OLTP monitors, DBMS, middleware, etc.)
Security	Security measures available.
Alternative Access	Front end applications, direct inserts allowed, application create transaction bypasses.
Other Applications	Interfaces to and dependencies on other applications.
Availability of data capture transactions	Manual or automated capture of data, replication of data from another application.

Evaluate Candidate Source for Future Plans

Assess plans for the data in terms of expansion, development, rationalisation or decommissioning.

4.1.2 Other Quality Properties of the Data Acquisition Component.

Due to the limited scope of this paper, only an outline is provided of the other Quality Properties of the Data Acquisition component. These quality properties are processes that contribute to data quality and include:

- Recast Data
- Edit Data

Recast Data

This is a *process* that is required for historical continuity where different data structures exist over time. This process is necessary to provide consistency where a data warehouse is used for trend reporting.

Edit Data

This is a *process* that provides checks on data entering the data warehouse, to ensure conformance to business rules and avoid low-quality data entering the data warehouse. This process includes the procedures to ensure that data that failed the edit check are passed back to the operational system for correction and reload.

4.1.2 Quality Contributors

The above processes within the Data Acquisition component of the data warehouse affect the following Data Quality Dimensions:

Quality Property of the Data Acquisition Component	**Data Quality Dimension**
Identify Best Source of Data	Correctness Currency Completeness
Recast Data	Consistency
Edit Data	Correctness

4.2 Quality Properties of the Data Storage Component

The following sub-section provides an outline of the quality properties of the Data Storage component of a data warehouse. These quality properties are processes or features of the Data Storage component that can be designed and implemented. This section shows the link between the quality properties of the Data Storage component and data quality dimensions.

Appropriate Data Structures

This quality property of the data storage component provides the database design of a data warehouse and ensures that the database includes data structures based on the objects and events that need to be reported on through the data warehouse. The database design will be based on dimension modelling (see [Kimball96]) and include all dimensions that are required for reporting.

Integrated Data

The quality property of integration is that data to be stored in the data warehouse is in an integrated, consistent form, despite being sourced from disparate data in the operational systems.

Integration in the database includes consistent naming conventions, consistent measurement of variables, consistent encoding structures, consistent physical attributes of data, and so on.

Time-Variant Data

This feature of the data warehouse database allows data to be included for a long time horizon (5-10 years) and the data elements are annotated with an indication of time.

Non-volatile Data

Non-volatility in the data storage component means that all data is accessed as read-only. This is in contrast to the operational systems which are volatile and are updated (inserts, deletes, changes) on a record-by-record basis. In the data warehouse data storage component, data is loaded once and no further updates are made.

Appropriate Levels of Granularity

This quality property provides that summaries of the base data be at the desired level. Appropriate levels of granularity includes ensuring that there is an appropriate level of detail to meet user requirements.

Correct Data Derivation Algorithms

This property ensured that the algorithms used for derived data are correct and will not provide erroneous results.

Quality Property of the Data Storage Component	**Data Quality Dimension**
Appropriate Data Structures	Consistency, Completeness, Accessibility
Integrated data	Consistency
Time Variant Data	Currency
Non-Volatile data	Consistency, Correctness
Appropriate Granularity	Accessibility
Correct Data Derivation Algorithms	Correctness

4.3 Quality Properties of the Information Catalogue

The following section provides an outline of the quality properties of the Information Catalogue component of a data warehouse. These quality properties are processes or features of the Information Catalogue component that can be designed and implemented. This section shows the link between the quality properties of an Information Catalogue component and data quality dimensions.

Data Dictionary

This property of the Information Catalogue contains formal data names and data definitions for all the data included in the data warehouse. It provides the base for understanding the real content and meaning of the data.

Description of Data Structure

This property is the recording of the structure of the logical data of the data warehouse data model.

Data Integrity Rules

This documents the rules for maintaining the data integrity of the data in the data warehouse.

Data Thesaurus

The data thesaurus contains a set of data name synonyms for the data in the data warehouse to help people locate the particular data they need.

Data Translation Schemes

Data Translation schemes provide a translation between common data characteristic variations that represent measurements or format variations.

Data Quality Profile

The Data Quality Profile provides a quality assessment that may be associated with each type of data. It provides the user with an indication of how reliable the information is.

The above properties affect the Data Quality dimensions as follows:

Quality Property of the Information Catalogue	**Data Quality Dimension**
Data Dictionary	Accessibility
Description of Data Structure	Accessibility
Data Integrity Rules	Correctness, Consistency, Accessibility
Data Thesaurus	Accessibility
Data Translation Schemes	Accessibility
Data Quality Profile	Correctness, Accessibility

4.4 Quality Model Framework for a Data Warehouse

The quality model framework for data quality in a data warehouse shows the link between observable data warehouse properties and the data quality dimensions. This is shown below in figure 5.

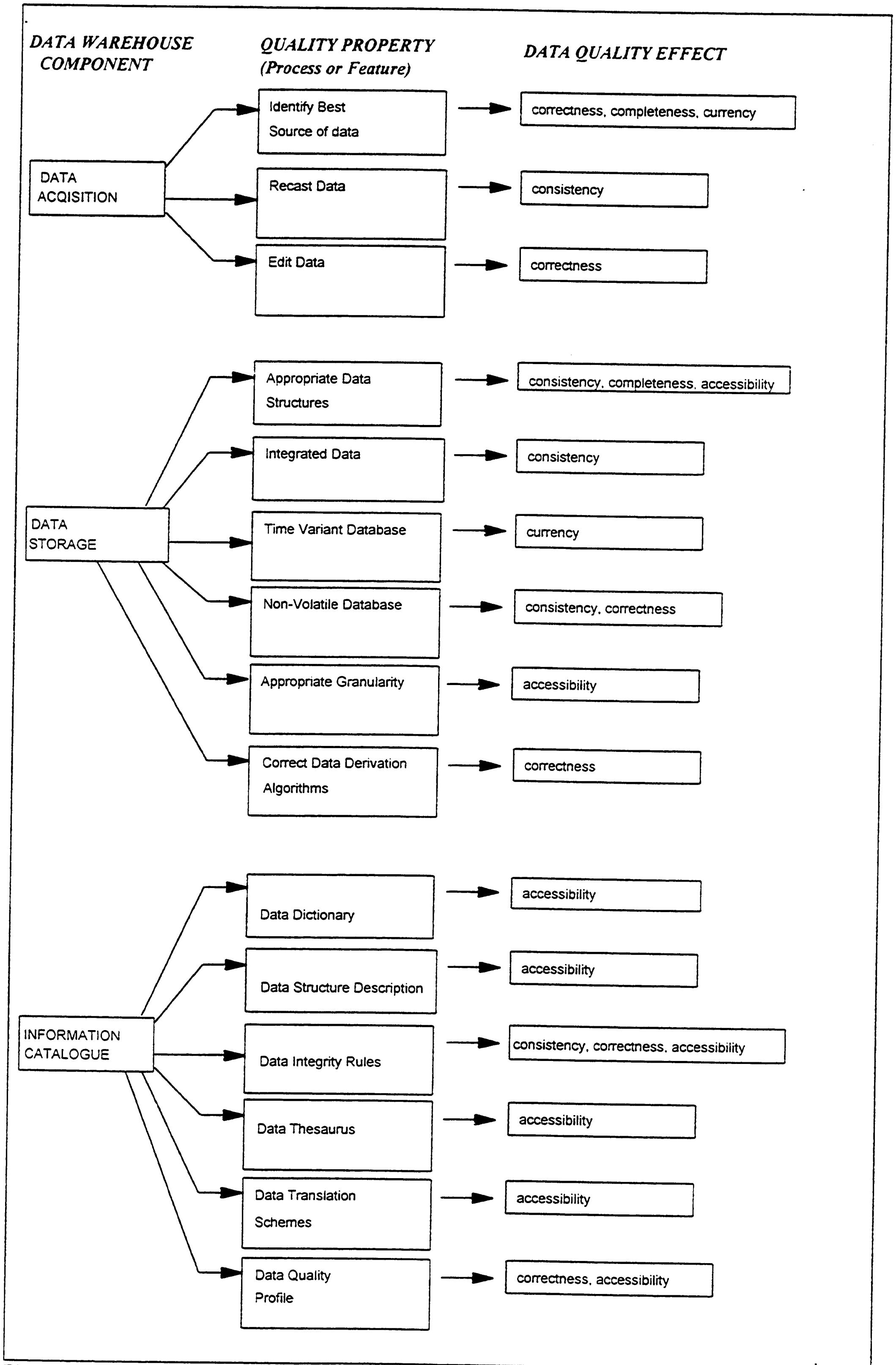

Figure 5. Data Quality Model Framework for a data warehouse

5. Conclusions

The main idea of this paper was to show that data quality can only be determined based on the intended usage and application of the data. This idea is different from most of the views expressed in the data quality literature. I have demonstrated that data quality depends on the intended use of the data through an example of a data warehouse which (usually) contains large volumes of data and is used to provide overall company reporting. A data warehouse reuses data that was originally captured for another purpose. I have shown in this paper that the same data held in both a data warehouse and in other systems can have different quality criteria, depending on how the data is intended to be used.

I have proposed five data quality dimensions to which properties of a data warehouse have been linked. The data quality dimensions I have chosen may not necessarily be the only ones worth considering but they are dimensions that other researchers have considered important. Further research can be undertaken to assess the applicability of other data quality dimensions for data in a data warehouse.

This paper provides a draft outline of a data quality handbook for a data warehouse. This data quality handbook identifies properties of a data warehouse that I believe have a direct bearing on the quality of data in the data warehouse. This data quality handbook is at an initial, draft stage and further work can be undertaken to provide the necessary detail.

References

Agmon, N., & Ahituv, N. Assessing data reliability in an information system. *Journal of Management Information Systems, 4(2), 34-44*, 1987.

Borenstein, N. *Programming as if People Mattered.* Princeton University Press. New Jersey, 1991.

Brackett, M.H. *The Data Warehouse Challenge: Taming Data Chaos.* Wiley Computer Publishing, 1996.

Cykana, P., Paul. A., Stern. M. DOD Guidelines on Data Quality Management. *Proceedings of the 1996 Conference on Information Quality.* 1996.

Deming E. W. *Out of the Crisis.* Center for Advanced Engineering Study, MIT, Cambridge, Mass. 1986.

Devlin, B.A. and Murphy, P.T. An Architecture for a Business and Information System. *IBM Systems Journal, Vol. 27, No. 1.* 1988.

Dromey, G. Cornering the Chimera. *IEEE Software. January* .1996.

Fox, C., Levitin. A., Redman. T. The Notion of Data and its Quality Dimensions. *Information Processing & Management. Vol. 30, No. 1, pp. 9-19.* 1994.

Heckel, P. *The Elements of Friendly Software Design.* Sybex. Alameda. 1991.

IEEE. IEEE Standard Glossary of Software Engineering Terminology. *IEEE Standard 729.* 1983.

Inmon, W. H. *Building the Data Warehouse.* John Wiley & Sons, Inc.1992.

Kent, W. *Data and Reality.* North Holland. New York, 1978.

Kimball, R. *The Data Warehouse Toolkit. Practical Techniques for Building Dimensional Data Warehouses.* John Wiley & Sons, Inc. 1996.

Mattison, R. *Data Warehousing Strategies, Technologies and Techniques.* Graw-Hill, 1996.

Mihulka, D. What is a Data Warehouse. *UNCNS Computing News.* 1995.

Norman, D. *The Psychology of Everyday Things.* Basic Books. New York. 1988.

Orman, L., Storey, V.C., & Wang, R.Y. Systems Approaches to Improving Data Quality. *Proceedings of the 1996 Conference on Information Quality.* 1996.

Percy, T. Data Warehousing - Passing Fancy or Strategic Imperative? *Gartner Group Conference presentation.* 1995.

Poe, V. Data Warehouse Architecture is not Infrastructure. *Database Programming & Design, March.* 1996.

Redman, T.C., *Data Quality: Management and Technology.* Bantam Books. New York, 1992.

Strong, D., Lee, Y. and Wang, R. Data Quality in Context. *Communications of the ACM, May.* 1997.

Wand Y. and Wang R.Y. Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM. 1996, Vol. 39, November. pp. 86-95.* 1996.

Wang, R. Y., Reddy, M.P., Kon, H.B. Data quality requirements analysis and modeling. *Proceedings of IEEE 9th International Conference on Data Engineering. Held Vienna, Austria. 19-23 April.* 1993.

Wang, R.Y., Storey,V.C., Firth,C. A Framework for Analysis of Data Quality Research. *IEEE Transactions on Knowledge and Data Engineering. Vol 7. pp. 623-40. August.* 1995.

Wang, R. and Strong, D. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information systems. 12. April.* 1996.