Developing Data Quality Metrics for a Product Master Data Model

THESIS

submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK INFORMATION ARCHITECTURE

by
Marhendra Lidiansa
4256360

Web Information Systems Research Group
Faculty EECMS, Delft University of Technology
Delft, The Netherlands
www.ewi.tudelft.nl

Elsevier B.V.
Strategy, Global Operations
Amsterdam, The Netherlands
www.elsevier.com

Contents

List of Tables

List of Figures

Executive Summary

Master data management (MDM) is implemented to increase the quality of core business data by having a single managed repository. Like any other IT projects, there are failures in the implementation of MDM. Several main causes of failures in MDM implementation are related to a missing data quality process, for example, a lack of proactive data quality surveillance (Sivola et al. [24]) and a lack of data quality measurements (Haug [10]). An important phase in the data quality process is the measurement phase that exercises the data quality metrics. In accordance with Elsevier's plan to implement product master data, the main objective of this study is **to identify, collect, analyze, and evaluate the quality metrics for a product master data; to allow quantifying and improve their value.**

In order to meet the main objective, this study needs to address these three questions: (1) What is the type of methodology that should be used to develop business-oriented data quality metrics? (2) How appropriate is the methodology for a practical case? (3) What are the data quality metric specifications for a case study in Elsevier? There are four phases in this thesis work to develop and answer those questions. In the first phase, the introduction phase, the main objective and research questions are formulated with several considerations, particularly the scientific and practical benefit of the study, and the boundaries of the projects.

In the second phase, the conceptualization, we need to select or construct the general process framework (GPF) to develop business-oriented data quality metrics as the answer to the first question. This study selects methodology developed by Otto [19] as the GPF to develop the data quality metrics. The selection process is conducted by comparing the methodology with other methodologies—like AIMQ, TIQM, DQ-P, and ORME-DQ—on several features, for example, the process model, metamodel, business needs consideration in its data quality criteria, and the focus of the method. Other studies in data quality (DQ) requirements, DQ metrics specification, DQ metrics requirements integration, data modelling, and process modelling are also used to ensure that the process model and metamodel in the selected GPF are adjustable for the case study.

The background of this thesis is related to the MDM system which function is to provide numerous enterprise applications with high quality critical business objects. Thus, we need to make sure that the developed data quality metrics meet the requirements of several business applications (Loshin [14]). This thesis uses the process model developed by Wang et al. [29] and Batini et al. [4] as the GPF to integrate data quality metrics from several applications into the product MDM repository. The activities include developing the appropriate data models, making the schemas conformed, and conflict resolution using qualitative criteria—completeness and correctness, minimality, and understandability. The result is a list of feasible data quality metrics that meet the needs of several applications

Here the thesis work uses the GPF as the first version of the developed solution to address the main objective. The main processes in the GPF are identifying the business problems and data defects, specifying the data quality requirement and metrics, verification of the result, and integrating the data quality metrics.

The thesis work addresses the third phase, the validation, by executing the GPF for a case study in Elsevier. Each process is adjusted with the case and analyzed for the required alterations. The activities

in this process consist of literature study and workshop/interview with the domain experts. The result of this activity is the altered GPF as the developed solution to address the main goal. The changes consist of an alternate configuration for the process model and the tangible objects for the components in the metamodel, for example, interview questionnaires, data quality requirements, data quality attributes, business problems–data defects matrix, and the data quality metrics. These results are used in formulating the answers to the second question. This thesis also conducts the testing activity by assessing the developed metrics with the criteria in the data quality metrics requirements.

The developed and filtered data quality metrics are feasible for the study case in Elsevier and include several data quality dimensions, namely completeness, syntactical correctness, absence of repetition, absence of contradiction, and accuracy. Those data quality metrics are the answers for the third question.

This thesis addresses its main objective by having two main results, namely a list of data quality metrics for eCommerce and product MDM system in Elsevier, and a process model to develop data quality metrics for a product MDM. The process model is developed on the basis of the works by Otto [19], Wang et al. [29] and Batini et al. [4], and considered practical, valid, complete, and resilience. This study also provides several lessons, for example, the critical success factors for each phase in the process model, recommendations for data quality process in Elsevier, and updates for product MDM data model. Furthermore, studies on the same issue with several other data/process domains are needed to get other possible configurations of the process model.

Thesis Committee:

Chair	: Prof. Dr. Ir. G. J. P. M. Houben, Faculty EEMCS, Delft University of Technology
Supervisor	: Dr. M. V. Dignum, Faculty TBM, Delft University of Technology
External Supervisor	: Olga Tchivikova, Director, Strategy, Global Operations, Elsevier

Acknowledgments

This document has been produced for the master thesis project as part of the Master of Science program in Information Architecture at TU Delft. The thesis work took place at Elsevier in Amsterdam from October 2013 until April 2014.

This thesis has been completed with the guidance, assistance, and support a number of people that I would like to thank.

First of all, my thesis supervisor at TU Delft, Dr. M. V. Dignum, with whom I've always discussed the thesis methodology, the thesis progress, the theoretical contents, and the practical approaches. Professor Dr. Ir. G. J. P. M. Houben, my professor, who provided necessary and critical comments at the initial phase of the project and during the midterm presentation. His inputs were important to develop the thesis goal and to make sure that the work was in the right direction.

My supervisors at Elsevier, Olga Tchivikova and James Carne, who were very open toward my thesis work and provided an association to their internal project. They also arranged all the resources I needed to complete my work and made sure other personnel provided the information I required. I would also like to thank the personnel in Elsevier for answering my countless questions, providing me with much-needed information and inputs, and helping me to develop the data quality metrics. Those people are the domain experts at Elsevier in e-commerce, marketing, book and journal data, data quality, IT operation, and IT infrastructure.

Elsevier provided a stimulating working environment, and it also has interesting problems closely related with the information architecture track, for example, data and text analytics, information retrieval, and recommender systems. I can recommend to anyone who is interested to conduct a thesis work at Elsevier.

1 Introduction

1.1 Background

Master data management (MDM) is a collection of the best data management practices that orchestrate key stakeholders, participants, and business clients in incorporating business applications, information management methods, and data management tools to implement policies, procedures, services, and infrastructures to support the capture, integration, and subsequent shared use of accurate, timely, consistent, and complete master data (Loshin [14]). According to DAMA International (DAMA [5]), the MDM has three goals: providing an authoritative source of high-quality master data ("golden record"), lowering cost and complexity through standards, and supporting business intelligence and information integration.

The importance of implementing MDM is gaining more prominence in companies. The Information Difference[1] reported that MDM projects are growing around 24% in 2012 (USD1.08 billion in the software market). This figure was also predicted by Gartner in 2011.

1.2 Problems

Companies incur cost when cleaning and ensuring high-quality master data (direct) and from faulty managerial decision making (indirect) caused by poor-quality master data (Haug [10]). The process failure costs due to bad-quality data comes from information scrap and reworking costs to improve the data quality (English [7]). Another cost caused by bad-quality data is missed opportunity costs.

While MDM is expected to lower those costs, the implementation of MDM could still provide master data with low data quality. Some barriers in achieving high-quality master data are lacking data quality measurements and lacking clear roles in the data life-cycle process (Haug [10]). The preconditions for a company in implementing MDM to answer the challenge of poor-quality data are (Sivola et al. [24]) a common definition of the data model to be used across the organization (data model), a proactive data quality surveillance (data quality), and a unified data model (information system).

Thus, data quality management is an important concern in MDM to provide high-quality master data (Loshin [14]), and the inclusion of data quality assessment and improvement activities is a key success factor to have a successful MDM project.

1.3 Focus and Demarcation

Data quality and MDM are broad subjects to study. There are several concerns that should be addressed in implementing an MDM, for example, stakeholders and participants' involvement, metadata management, architecture styles, functional services, data governance, data modeling, data consolidation and integration, management guidance, data quality management, master data identification, and master data synchronization (Batini et al. [4]; Loshin [14]). There are also several data domains in MDM, for example, customer, product, supplier, material, and asset (Otto [18]), where each can serve several business processes in an organization.

[1] http://www.informationdifference.com/products/landscape/mdm-landscape/index.html

The phases within the data quality assessment and improvement activities (Batini et al. [3]) that can also be used within an MDM are the following:

i. State Reconstruction
   The aim of this phase is to get information about business processes and services, data collection, quality issues, and corresponding costs.
ii. Assessment/Measurement
   The aim of this phase is to measure the quality of data collection along relevant quality dimensions. The results from the measurement activity are further compared with certain reference values to determine the state of quality and to assess the causes of poor data. Defining the qualities, dimensions, and metrics to assess data is a critical activity.
iii. Improvement
   The aim of this phase is to select the steps, strategies, and techniques that meet the new data quality targets.

To limit the scope, in order to achieve a good and reasonable goal within a limited given time and resource, this thesis will focus on these parts:

i. Data Quality Phase
   The first phase, state reconstruction, is considered optional, and some methodologies only use existing documentation to develop information of the business process and information system (Batini et al. [3]). The data quality metrics and dimensions are important entities in the second phase of data quality improvement methods. Those entities are discussed in all 13 methods assessed by Batini et al. [3]. The improvement phase is found in 9 of 13 methods, and it covers 10 activities. The third phase is considered more extensive than the second phase because it also covers the business process (e.g., process redesign) and organizational aspects (e.g., assignment of data stewardship responsibilities and data quality improvement management).
   Thus, this study will focus on the assessment/measurement phases, especially on the development of qualities, dimensions, and metrics. Another reason is because the assessment phase result determines the courses of action in the improvement phase.
ii. Data Domain
   Dreibelbis et al. [6] classify the master data domain into three categories, namely party, product, and account. Most of the MDM software products serve the customer (party) and the product domains for the same reasons, for example, those entities are important for the business, and those entities are used to identify them from their competitors. Elsevier has developed a customer (party) MDM, and it is initiating the product MDM implementation.
   This study will focus on product data so the results can be used as design artifacts for the product MDM implementation, particularly in Elsevier.

1.4 Research Goal and Questions

The main research goal of this study is **to identify, collect, analyze, and evaluate the quality metrics for a product master data; to allow quantifying and improve their value**. The identification of data quality metrics should be for the ones that provide business impacts for the organization. This requirement is relevant for an MDM system, a repository for important business objects. According to

English [6], pragmatic information quality is the value that accurate data has in supporting the work of the enterprise i.e., data that does not help enable the enterprise accomplish its mission has no quality.

In order to address the main objective, several research questions are constructed, which are as follows:

i. What is the type of methodology that should be used to develop business-oriented data quality metrics?
   Strategy: Study the literatures on several data quality assessment and improvement methods. Analyze the main goal, the process model, and the metamodel defined in each study.
   Objective: Select an appropriate metamodel/process model that best fits the main research goal. The selected metamodel/process model will be the base reference for the thesis work's activity.

ii. How appropriate is the methodology for a practical case? What processes or components should be altered?
   Strategy: Conduct the activities to identify, collect, analyze, and evaluate the quality metrics for a product data in the Elsevier environment using the selected metamodel/process model. Analyze the process, result, and findings to assess the compatibility of the methodology with the study case in Elsevier.
   Objective: This is a theory-testing strategy, and its aim is to improve or adjust the selected metamodel/process model on the basis of the findings (if any) to fill in the missing factors that were not determined within the scientific area of this topic (Vershcuren [27]).

iii. What are the data quality metrics specifications for a study case in Elsevier?
   Strategy: Conduct the activities to identify, collect, analyze, and evaluate the quality metrics for a product data in the Elsevier environment using the selected metamodel/process model. Assess the acceptability of each metrics for each requirement; for example, conduct a data quality and performance assessment for each metrics in the data repository to assess its feasibility.
   Objective: Provide a list of relevant and acceptable DQ metrics for the case study in Elsevier.

1.5 Research Approach

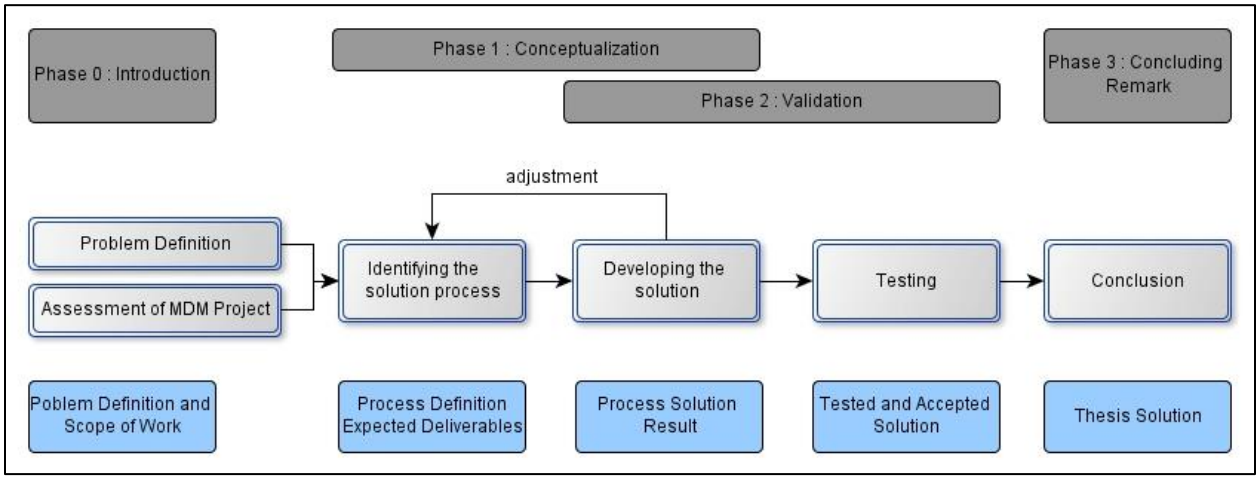


Figure 1 Research Approach

This study will have several activities in Figure 1 to answer the research goal. The research is composed of four main phases, namely the introduction, conceptualization, validation, and concluding remarks. The descriptions of the activities within those phases are as follows:

i. Problem Definition and Assessment of the Product MDM Project

The activities within this phase have a goal to define a specific problem in the MDM implementation, the scope of the thesis work, the goal for the thesis. The activities consist of literature study of the MDM and data quality in general to understand the components and critical information needed in the MDM project. That information is used to assess the product MDM project in Elsevier, which was conducted in March–August 2013, in order to identify this information: the goal of the project and the activities that have been conducted, the critical components that are expected to develop but have not been delivered, and the schedule for the project plan.

Important results of this activity are the research goal in section 1.4, the cases in Elsevier that are relevant to the thesis goal and the deliverables of the study.

The results of the assessment are

a. The results of his thesis, —that is, the data quality metrics and the process model—, could be used in the Elsevier's MDM project.
b. There are two cases that will be used in this thesis work:
   - Elsevier e-commerce
     This case is selected because e-commerce is one of the main product data consumers and its end customers are the web users who buy the book or journal. The web users use product information for their buying decisions. The data quality specification for the product data should be developed to be useful and usable for the information consumers (Lee [13]) and consistently meet customers' expectations (English [7]).
   - Elsevier Customer System (CS)
     This case is selected because this system provides the books and journals metadata for a system in Elsevier that also functions as an e-commerce. The customers use the metadata provided by this system to purchase a book or a journal.
c. To align with Elsevier project, this thesis selects product master data for the master data object with a specific domain in book and journal data.

ii. Identifying the Solution Process

There are two main activities in this phase, which main strategy is using available literature studies. First, this thesis selects the study that provides a process model and metamodel to address the main objective. Second, this thesis studies other studies to enrich the components in the selected metamodel.

a. Literature study to define the general process framework (GPF)
   A study on several methodologies to develop data quality metrics that consider the business impacts of having poor data quality is conducted to select one as the base process framework. This activity is needed to provide a structured and proven method for the research and conducted before developing the solution for a study case in Elsevier.
b. Literature study to define the components within the GPF

The selected general framework could also provide tangible objects for each component that will be used within its activities, for example, the definition of data quality and the metrics. This study will consider other studies results to provide options for the activities and selected components for these reasons:

- There could be more updated studies on the components from other studies that can be attached in the process framework.
- To be more flexible in developing the solution for the case studies in Elsevier because of their unique situation, for example, the degree of document completeness and the degree of process complexity.

This activity is conducted before and during the development of the solution in Elsevier's case. The result of this activity will provide the answer for the first research question: *what is the type of methodology that should be used to develop business-oriented data quality metrics?*

iii. Developing the Solution

Here, the thesis conducts the theory testing research where the aim is to test and make adjustments if necessary (Verschuren et al. [27]) to the GPF defined in the previous step. The process solution proposed in previous stage is then implemented in Elsevier's environment once the GPF is selected. Some details of previous projects are also used in this step with user's validation to avoid recurring activities.

This step will be conducted simultaneously with step II because the situation in Elsevier needs to be identified at an earlier phase. This enables us to make necessary adjustments for the GPF. An example is the required adjustment if the process metrics (e.g., KPIs) for Elsevier's business process are not available.

Interaction with the Elsevier system and experts is required to assess the business process and information system, to identify the quality metrics for the product MDM data, and to validate the results. The adjustments for the process framework and the assumptions used to select a component will be documented as the research's result.

iv. Testing the Result

The test is needed to ensure that the quality metrics are related and useful to business performance. The qualitative reasoning that describes the relationship between the data quality metrics and business performance has been developed in the previous step. However, we also need to assess its compliance with other data quality metrics requirements. An example is we need to conduct a data quality assessment process to some applications to assess the degree of feasibility and reproducible. The result of activity III and IV will provide the answers for the second and third research questions as follows:

- *How appropriate is the methodology for a practical case? What processes or components should be altered?*
- *What are the data quality metrics specifications for a study case in Elsevier?*

v. Concluding Remark

This final phase is not a project execution. It describes the conclusions, lessons learned, and the future recommendation on the basis of the findings in phase2 validation.

1.6 Structure of the Thesis

The process, results, and findings during the project will be described in the thesis document with the following outline:

i. Introduction

There are two chapters in this section as follows:

a) Chapter 1 - Introduction

This chapter introduces the background and the problems in the MDM environment that require the data quality assessment and improvement activity. Some limitations are introduced to develop the boundaries of the thesis and to make sure that the work is within a master thesis project's load. The research's goal and questions are set on the basis of the defined background, problems, and boundaries.

The research approach is developed as a guide to conduct the thesis work. The activities in the research approach reflect the structure of the thesis, which includes introduction, conceptualization, empirical evaluation/ validation, and conclusion.

b) Chapter 2 - Project Environment

This chapter describes the working environment when conducting the thesis work in Elsevier with some explanations of the organization, information system, and resources.

ii. Conceptualization

There is one chapter with two subchapters in this section as follows:

a) Chapter 3.1 - Related Studies

This chapter describes several studies that are related to the thesis project and needed to answer some of the research questions to attain the research goal. The studies are mostly for subjects in data quality and MDM, the main topic of this thesis project.

b) Chapter 3.2 - General Process Framework

To answer the research questions, this study needs to have a more concrete and scientifically sound process model and metamodel. The process model and metamodel is selected from other studies, and it will be used as a base reference method for the validation phase.

iii. Validation

The chapter in this section, Chapter 4 – "Empirical Evaluation in Elsevier," describes the process and results when conducting the GPF in the case study environment. It also describes the findings when conducting the GPF in the case study environment. The findings are the alteration for the process model to be feasible for the case study.

iv. Concluding Remarks

The chapter in this section, Chapter 5 – "Conclusion," provides the summary of the thesis works and the expected future works.

2 Project Environment

2.1 Elsevier

Elsevier is the world's leading provider of scientific, technical, and medical (STM) information and serves more than 30 million scientists, students, and health and information professionals worldwide. It has over 11 million articles and over 11,000 full-text e-books in one of its platforms, ScienceDirect. According to its 2012 financial report, Elsevier provided more electronic products. A total of 64% of the revenue is from electronic format (21% is print/other), and 85% of titles in STM are available in electronic format.

As described in section 1.3 and 1.5, the focus of the study is on book/journal product data in the e-commerce environment. However, we need to reduce the business domain in Elsevier because it has several Web commerce platforms to sell the products. Each platform could have different responsible units, e-commerce systems, target customers, and product types. Thus, we limit business domain for the direct marketing under the e-Business division. This provides the description of e-commerce as follows:

a. Organization

   Product, Marketing, and Sales unit under the e-Business division

b. E-commerce platform

   The e-commerce platform (Figure 2) is using the e-commerce system as the main platform, and it is limited to several websites, namely e-store (store.elsevier.com; B2C), B2B site, and other e-commerce websites.

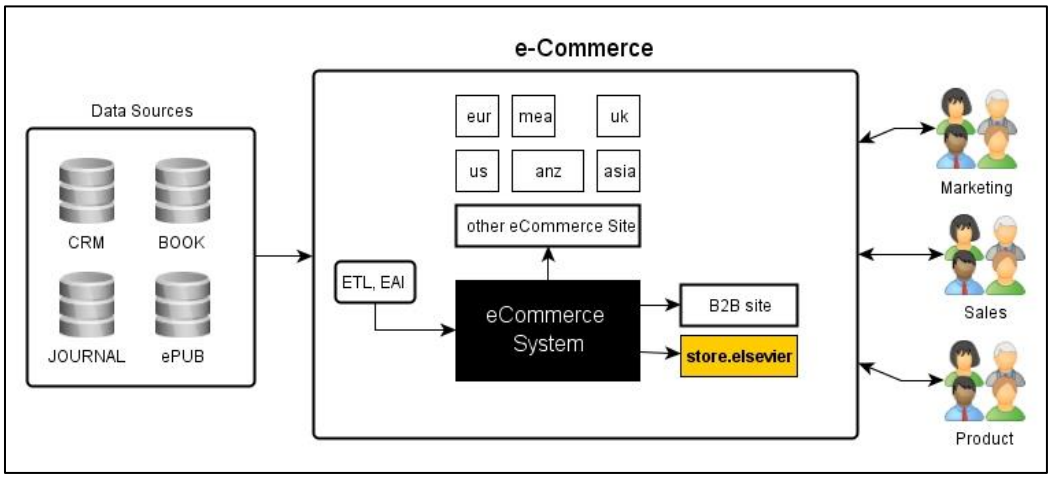


Figure 2 E-commerce in Elsevier

2.2 Resource

During the thesis work, Elsevier provides the graduate the opportunity to use several physical and nonphysical resources as follows:

i. A mentor to guide the thesis project, providing the right resources and people to interview, and to distribute the information requirements.

ii. Domain experts who provide important information for the study through interview and document sharing.

iii. An access to ScienceDirect.com from the office to get required electronic research papers, journals, and books.
iv. Notebook, e-mail, and access to internal network. These resources allow me to access the internal application useful for the research and to contact internal staff to ask information related to the study.
v. Knowledge repository to access related documents, for example, project documents, operational documents, and architecture documents. There are two main resources used for this study, namely CBS Wiki and MOSS document management system.
vi. A collaboration application to conduct online meetings.

3 Conceptualization

To answer the first research question, -*what is the type of methodology that should be used to develop business-oriented data quality metrics?*-, we need to review existing studies in data quality. There are two components that this study needs to establish, namely the process model and the metamodel. The process model is required to provide the activities, description and method of the activities, goal of each activity, and their sequence to develop the data quality metrics. The metamodel model is needed to provide the components, definition, relationship among the components, and how it is used by the activity in the process model. Those two main components provide the general process framework (GPF) that will be used as the base reference for this study.

3.1 Related Studies

3.1.1 Data Quality Methodology

3.1.1.1 AIMQ

AIMQ (Lee et al. [13]) is a methodology for information quality (IQ) assessment and benchmarking. The methodology is developed on the basis of other academic studies (e.g., Wang & Strong, Goodhue, Jarke & Vassiliou) and several practitioners' view (e.g., Department of Defense, HSBC, and AT&T), and is validated using cases from three large health organizations. The methodology consists of a model of IQ, a questionnaire to measure IQ, and analysis techniques in interpreting IQ.

The important components in AIMQ are the IQ model and IQ dimensions, which are critical for the information consumers. The IQ model in AIMQ, PSP/IQ model, has four quadrants that are relevant to an IQ improvement decision. Those four quadrants are sound information, useful information, dependable information, and usable information.

This model is used to assess how well an organization develops sound and useful information products and delivers dependable and usable information services to the consumers.

3.1.1.2 Total Information Quality Management (TIQM)

English [7] defined quality as consistently meeting customers' expectations. TIQM (Figure 3) is developed on the basis of quality management principles, techniques, and processes from the leaders of the quality management revolution and has the following processes related to data quality assessment:

a. P1 Assess Data Definition and Information Quality Architecture
   This process defines how to measure the quality of data definition to meet the knowledge workers' requirements, current information architecture and database design quality, and customer satisfaction with data definition.
b. P2 Assess Information Quality
   This process defines how to measure the quality of information to meet the various quality characteristics, such as accuracy and completeness. Activities related to this process are identifying information quality objectives and measure, identifying data and reference data, measuring the information quality, and reporting information quality.

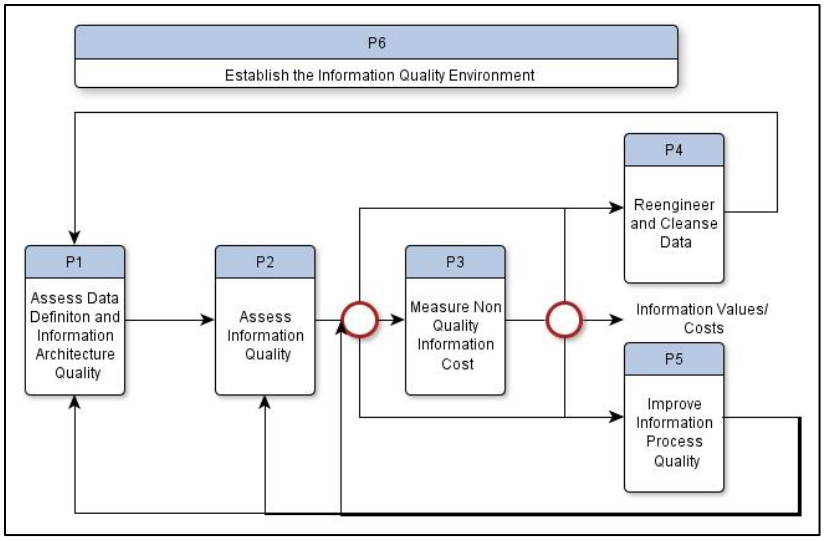


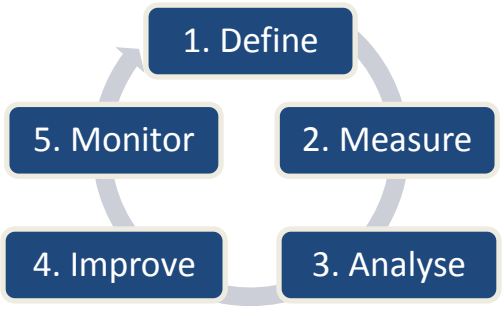
Figure 3 TIQM Process, English

3.1.1.3 DQ-P

Morbey [17] defined data quality as the degree of fulfillment of all those requirements defined for data, which is needed for specific process. His study provides 7+2 data quality dimensions and examples of measurements. It also provides a general data quality process (DQ-P) with the following activities (Figure 4):

a. Define the data quality requirements, including the business rules, quality dimensions, and metrics
b. Measure the data in repository against the data quality requirement
c. Analyze the measurement result to identify the defects and the target for improvement
d. Improve the data quality by fixing the data or implementing new plausible checks at data input for future prevention
e. Monitor

Figure 4 DQ-P Activities, Morbey [17]



3.1.1.4 ORME-DQ

Batini et al. [2] provided several dimensions that can be used in data quality assessment, and they also provided a classification of costs and benefits that can be used to support decision in engaging data quality improvement campaigns. In general, they classified the costs into three categories, namely (i) the costs of current poor data quality, (ii) the costs of DQ initiatives to improve it, and (iii) the benefits that are gained from such initiatives. The benefits are also classified into three categories, namely (i)

monetizable, (ii) quantifiable, and (iii) intangible. They provided a data quality method, the ORME-DQ, which has these core steps:

a. Phase I: DQ Risk Prioritization
   Assessing and modelling relevant databases, business processes, potential loss of poor data, and correlation matrix
b. Phase II: DQ Risk Identification
   Evaluating economic loss to select critical processes, datasets, and data flow
c. Phase III: DQ Risk Measurement
   Conducting qualitative and quantitative assessment of data quality in current data repositories
d. Phase IV: DQ Risk Monitoring
   Evaluating the DQ dimension values periodically and sending alert when less than predefined values

3.1.1.5 Hybrid Approach

Woodal et al. [29] defined data quality as fit for use. They studied eight data quality assessment and improvement methodologies to provide a hybrid approach with recommended activities as follows: (a) select data items, (b) select a place where data is to be measured, (c) identify reference data, (d) identify DQ dimensions, (e) identify DQ metrics, (f) conduct measurement, and (g) conduct analysis of the results. The assessed methodologies include AIMQ (Lee et al. [13]), TQDM (English, 1999), cost-effect of low data quality (Loshin, 2004), and subjective-objective data quality assessment (McGilvray, 2008).

The conduct measurement activity (f) obtains values for the dimensions (d) and metrics (e) for a given set of data items (a). The measurement process applies the metrics to the data in a certain data repository (b). The process could be using reference data (c) depending on the type of datasets.

The methodology is validated using cases in a UK car manufacturer organization and the London Underground. An important result of this study is that the process model is configurable, and we could develop an alternate configuration of the process model for a certain domain or case study.

3.1.1.6 Business Oriented Data Quality Metrics

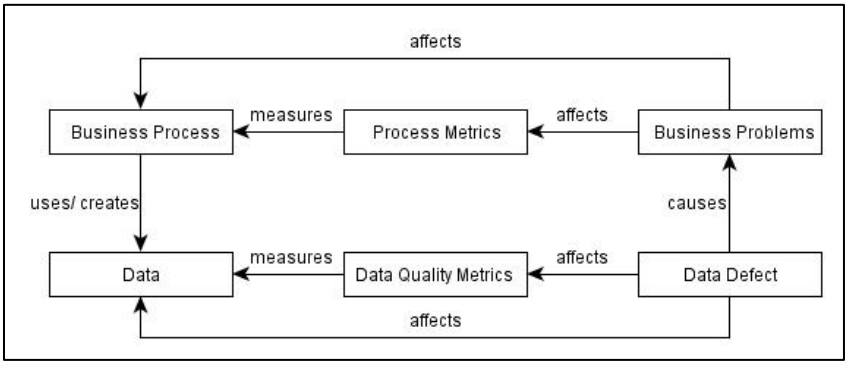


Figure 5 Entities and Relations of a Business-Oriented Data Quality Metric, Otto et al. [19]

Otto et al. [19] developed a methodology to identify business-oriented data quality metrics on the basis of 26 previous studies, among them are the studies by Batini (2006, 2007, 2009), Lee et al. (2002, 2006), English (1999), DAMA (2009), Loshin (2001), and Wang (1996). The methodology is developed

with the assumption that data defects could cause business problems (Figure 5) and that the identification of data quality metrics should be based on how the poor data impacts process metrics.

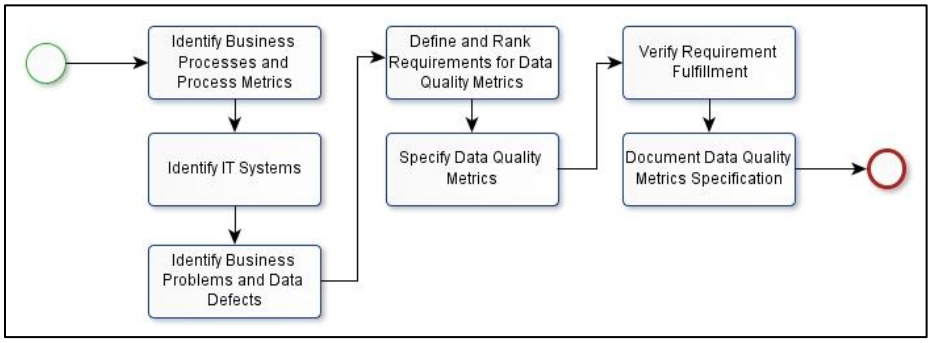


Figure 6 Process Model for DQ Metrics Identification, Otto et al. [19]

This study compared the methodologies to develop a process model with these considerations:

i. The structure of the method that consists of activities (Figure 6), results, metamodel, role, and techniques.

   The process model provides the activities consisting of three phases—namely, the identification phase to identify business problems and data defects, the specification phase to develop the data quality requirements and data quality metrics, and the verification phase to ensure the data quality metrics meet the requirements.

   The metamodel provides the components that should be developed within each activity in the process model. It also defines the relationships among the components in Figure 5.

ii. Pragmatic definition and representation.

iii. Relationship with business performance.

iv. Implementation capability.

This methodology is validated using cases in a telecommunication provider company (customer service processes), a watch movement producer company (manufacturing processes), and a machine manufacturer/automotive industry supplier company (maintenance processes).

3.1.2 Data Quality Metrics Requirements

There are studies in data quality that provide a number of data quality dimensions, measurement methods, and scales for the output value. Data quality metrics requirements are needed as a guide to define the data quality metric where a possible activity is to select for available lists. Data Management International provides several data quality dimensions—namely, accuracy, completeness, consistency, currency, precision, privacy, reasonableness, referential integrity, timeliness, uniqueness, and validity within its Data Management Body of Knowledge (DMBOK). DMBOK, which is authored by data quality practitioners, provides general requirements for the DQ metrics as in Table 23. There are also several studies that provide more specific requirements for DQ metrics:

a. Heinrich et al. [11] provided requirements on metrics value and scale (comparable, aggregation, interval) and methods (aggregation) to use in Table 24. Those requirements are developed from the study by Even and Shankaranarayanan (2007).

b. Reeve [25] provided generic requirements for effective measurement of data quality metrics

c. Hüner et al. [12] provided generic and specific requirements for DQ metrics in Table 30. This study developed the requirements using the study by DAMA [5] and Heinrich et al. [11]. The study also added some requirements—namely, understandability and complete information and relation with other components.
d. Loshin [15] provided generic and specific requirements for DQ metrics in Table 27. Some of the requirements are used by DAMA [5].

3.1.3 Data Quality Metrics Integration

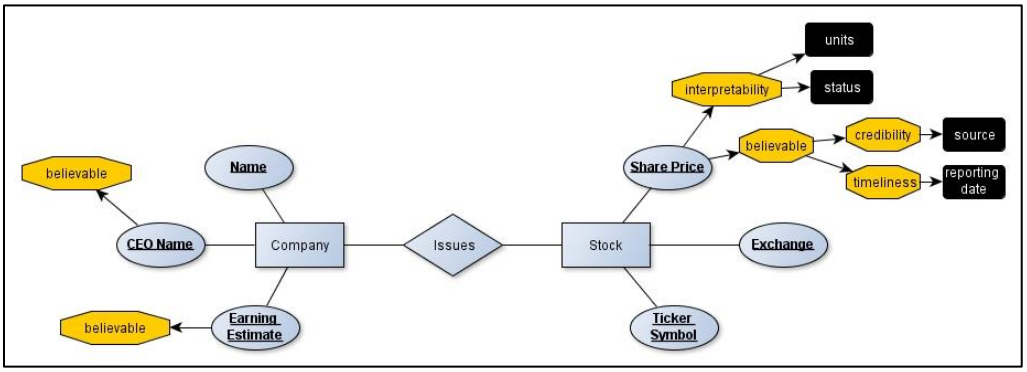


Figure 7 Data Model with Quality (Attribute Level), Wang et al. [29]

Wang et al. [29] provided the representation of quality as additional information within a data model (Figure 7). They also acknowledged the possibility of developing the data model with its quality requirements from several different applications, such as finance and human resource. The steps to acquire the data model with quality attributes by Wang et al. [29] are as follows (Figure 8):

1. Determine the application view of the data.
   The architect should develop a conceptual data model (ER diagram or class diagram) that is derived from existing application or business requirements. The result of this activity is the application view, an ER diagram of an application (Figure 7, blue shapes).
2. Determine (subjective) quality parameters for the application.
   The business users should determine the quality parameters to make the information accurate; for example, the CEO name should be believable or share price should by interpretable to be usable. The result of this activity is the parameter view, an application view with data quality dimensions for each attribute (Figure 7, yellow shapes).
3. Determine (objective) quality indicators for the application.
   Together with data architect, business users could define the objective quality, for example, the unit for share price and the trusted data sources. The result of this activity is the quality view, a parameter view with data quality indicators for each dimension (Figure 7, black shapes).
4. Conduct quality view integration.
   A data/information could be used by several applications, and each application could have different representations or requirements. This step should integrate from several views and be agreed by the business users, and this step is required to make sure that a variety of data quality requirements can be met. The consolidation process is considered similar to schema integration identified by Batini et al. [4].

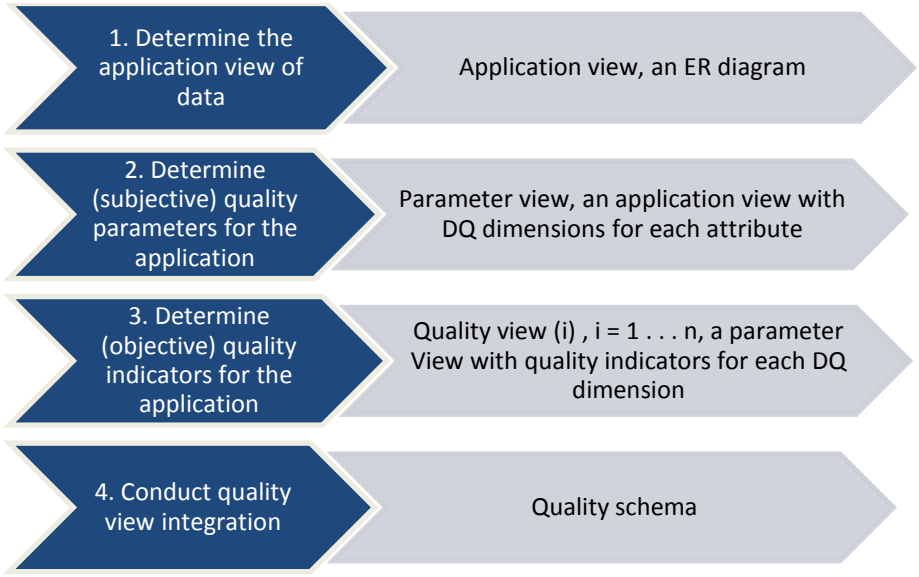


Figure 8 Steps of Identifying Quality Attributes, Wang et al. [29]

Batini et al. [4] provided several activities to integrate the schemas as follows:

a. Pre-integration
   An analysis of schemas is carried out to provide the information needed to make these decisions: the number of schemas to be integrated, the amount of designer interaction, the order of integration, and a possible assignment of preferences to entire schemas or portions of schemas.
b. Comparison of the Schema
   The schemas are compared to determine the correspondence among concepts and possible conflicts.
c. Conforming the Schema
   The goal of this activity is to resolve the schema conflicts with the designers and users before merging them.
d. Merging and Structuring
   The consideration for merging activity is using these qualitative criteria:
   i. Completeness and Correctness
      The integrated schema must contain all concepts present in any component schema correctly.
   ii. Minimality
      A concept must be represented once in the integrated schema if it is represented in more than one component schema.
   iii. Understandability
      Among several possible results, the one that is (qualitatively) the most understandable should be chosen.

3.1.4 Data Quality Dimensions

"The quality of data depends on the design and production processes involved in generating the data. To design for better quality, it is necessary first to understand **what quality means** and **how it is measured**" [28].

3.1.4.1 Quality dimensions by Sebastian-Coleman

Sebastian-Coleman [23] developed the data quality dimension as a component of DQAF measurement method. The quality dimensions in Table 15 are developed using these considerations: DQAF is used to define objective measures; DQAF is used for overall data management including basic controls that confirm receipt of data, measure the efficiency of technical processes in the data chain, and measure the quality of data content; and DQAF is used for in-line measurements. The study included several dimensions, such as completeness, validity, consistency, and integrity.

3.1.4.2 Quality dimensions by Morbey

Morbey [17] provided only 7+2 dimensions (7 automatically measurable and 2 documentary) for data quality with assumptions that other quality dimensions should already be checked by other teams in the company, namely expert approval, surveys, IT security/business monitoring, automatic measuring, visual inspection or document check, and audits/follow-up examinations. The essential dimensions of data quality are as in Table 16 which consists of completeness per row, syntactical correctness, absence of contradiction/ consistency, business referential integrity, absence of repetition/ uniqueness, and accuracy.

3.1.4.3 Quality dimensions by Batini et al.

Batini et al. [3] studied 13 methodologies of data quality assessment and improvement in 2009 and provided several basic sets of data quality dimensions, including accuracy, completeness, consistency, and timeliness. However, a general agreement on which set of dimensions defines the quality of data or on the exact definition of each dimension is not available. Batini et al. [3] defined quality dimensions as in Table 17. Several findings on data quality by Batini et al. [1] are as follows:

- Data quality is a multifaceted concept, as in whose definition different dimensions concur.
- The quality dimensions, such as accuracy, can be easily detected in some cases (e.g., misspellings) but are more difficult to detect in other cases (e.g., where admissible but not correct values are provided).
- A simple example of a completeness error has been shown, but as to accuracy, completeness can also be very difficult to evaluate, for example, if a tuple representing a movie is entirely missing from the relation movie.
- Consistency detection does not always localize the errors.

3.1.5 Data Quality Measurements

Data quality is a multidimensional concept, and companies must deal with both the subjective perceptions and the objective measurements on the basis of the dataset in question (Pipino et al. [22]).

3.1.5.1 Pipino et al.

Pipino et al. [22] did not develop a specific measurement method for each dimension, but they provided several generic operations that could be used within an objective measurement as follows:

a. Simple Ratio
   The simple ratio measures the ratio of desired outcomes to total outcomes. Several dimensions could use this form like free of error, completeness, and consistency free of error, completeness, and consistency.
b. Min or Max Operation
   The minimum operation can be used for believability, and the appropriate amount and the maximum operation can be used for timeliness and accessibility. The minimum operator is used when the indicators have value in the permissible range. The maximum operation is used when a liberal interpretation is warranted, but we want to make sure that the value is within a permissible range.
c. Weighted Average
   A weighted average is an alternative to the minimum operator. This study indicates the use of this form only when the company understands the importance of each indicator to the overall evaluation of a dimension.

3.1.5.2 Batini et al.

Batini et al. [3] found that there are several measurements/metrics on a single dimension on the basis of assessment to 13 methodologies of data quality improvement as in Table 19. The study provided the subjective and objective measurement methods as defined in the researched methodologies. An example of subjective measurement is by having a survey to the data consumers to assess the data quality level or their level of satisfaction. While the example for the objective measurement is by defining the criteria for a certain data quality attribute and developing the appropriate mathematical function for assessment.

3.1.5.3 Peralta

Peralta [21] studied only the accuracy and the freshness data quality dimension. The study provided three types of metrics that are used for the accuracy dimension, as follows:

a. Boolean metric: It is a Boolean value (1=true, 0=false) that indicates whether a data item is accurate (correct, precise) or not.
b. Degree metric: It is a degree that captures the impression or confidence of how accurate the data is. Such degree is commonly represented in the [0–1] range.
c. Value-deviation metric: It is a numeric value that captures the distance between a system data item and a reference one (e.g., its real-world correspondent entity). Such distance is generally normalized to the [0–1] range.

3.1.5.4 Sebastian-Coleman

Sebastian-Coleman [23] provided measurement methods for several dimensions as in Table 20. The study defines the measurement for a number of data quality attributes, namely completeness, validity, consistency, integrity, and timeliness. It also determines when to execute the measurement on the basis

of the data criticality. The in-line measurements are conducted when the data enters the system, and they are for critical data, whereas the periodic measurements could be performed weekly or monthly, and they are for less critical data.

Table 1 Summary of DQ Measurement, Sebastian-Coleman [23]

DQ Attributes	In-Line	Periodic	Process Control
Completeness	v	v	v
Validity	v	v	-
Consistency	v	v	v
Cross-table integrity	v	v	-
Timeliness	v	-	-

3.1.6 Types of Data Model in MDM

MDM is a cycle of consolidation, integration, and synchronization, and there could be a specific data model at each phase in the cycle (Figure 9).

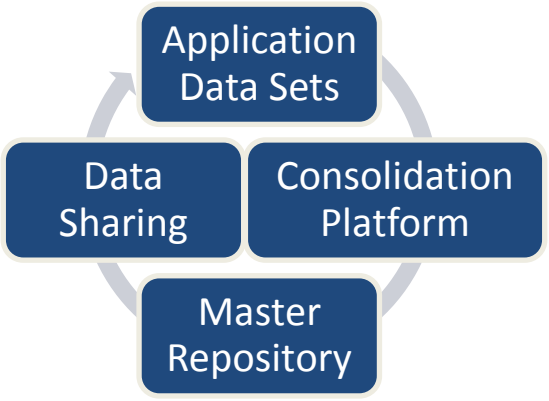


Figure 9 MDM Data Exchange, Loshin [14]

There are three data models that need to be defined in MDM, where one model is optional [14], as follows (Figure 10):

1. Exchange model for the exchange view
   This model is used to exchange the data between master repository and participant applications. The model captures the structure and data type of all participants' data that will form the master data object.
2. Consolidation model
   This data model is optional and acts as an intermediate model to consolidate the data sources into a master repository. Because the function is for consolidation, the object identifier definition is important in this model.
3. Repository model for the persistent view
   This model is used in the master repository as the model for master data object. This is the model that is viewed as the master data model by participant/target applications. Identity resolution is the challenge in this model because the object should be unique.

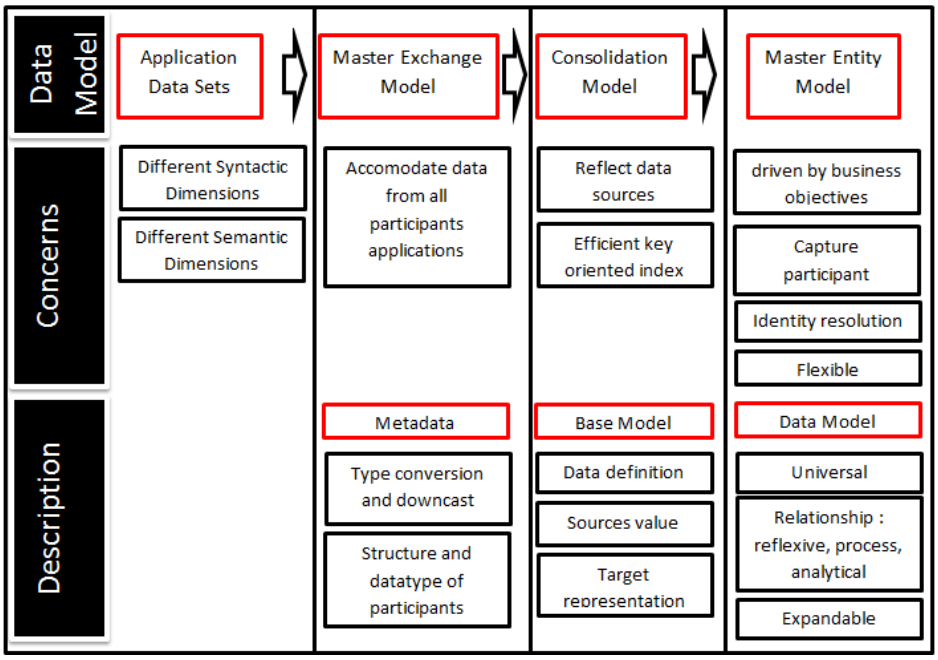


Figure 10 Data Model for MDM

3.1.7 Methodologies Comparison and Relationship with the Studies

As described earlier in 1.5, several studies on data quality methodology are needed to define the general framework. The goal is to find a methodology that provides clear procedures and required components in developing business-oriented data quality metrics. Studies in data quality methodology in 3.1.1 could be summarized in Table 2. This study compares the methodologies with the following criteria:

i. Process model

The initial phase of this research is to develop the general process framework (GPF) that provides complete and clear definition of its components. The existence of a process model and metamodel in the methodology is important because it will be used as a reference for the project to address the goal.

ii. Criteria for DQ

The goal of this thesis is **to identify, collect, analyze, and evaluate the quality metrics for a product master data; to allow quantifying and improve their value**. The objects in the master data are the high value ones and have business impacts for the company. Thus, it is important to understand how they define the criteria for data quality and whether they provide a correlation with the business needs.

iii. Scope of the Study

This thesis needs to focus on the measurement phase of the data quality assessment and improvement process. Each methodology could discuss a different set of phases and focus on certain parts/components in the metamodel.

Table 2 DQ Methodologies

Methodology	Process Model	Criteria for DQ	Scope
AIMQ (Lee et al. [13])	▪ Yes. It provides only DQ assessment on the basis of questionnaires and statistical function. ▪ The identification of task needs and standards is unavailable. The identification is more on how much data meet the standard or expectation by filling a value within a range.	DQ meets the standards and information consumer task needs.	Assessment/ measurement
TIQM (English)	▪ Yes. It is an ongoing process of reducing data defects. ▪ There is an identification of data definition quality and task requirements (completeness, accuracy, currency), but they are not related with possible business problems.	DQ meets the knowledge workers' requirements and standards.	Assessment/ measurement, improvement
DQ-P (Morbey [17])	▪ Yes. It has a generic process, like to define, measure, and analyze. ▪ Identification of task requirement is assumed done prior to the process. The DQ team starts its process after accepting the request for a DQ check.	DQ meets task requirements.	Assessment/ measurement, improvement
ORME-DQ (Batini et al. [2])	▪ Yes. It has a generic process and provides details on costs and benefits. ▪ Risk identification is important and conducted by assessing the cost-benefit.	Provide relation with potential loss of poor data.	Assessment/ measurement, improvement
Hybrid (Woodal et al. [30])	▪ Yes. Generic process with flexible activities and flow. ▪ Identification of task requirement is assumed done prior to the process.	Fit for use, i.e., meets task requirements.	Assessment/ measurement
Otto et al. [19]	▪ Yes. Only focus on metrics development and provide more details on activities. ▪ Identification of business problems and the cause of data defects is part of the process.	Provide relation to business problems, i.e., minimizes business problems.	Assessment/ measurement

The assessment of several methodologies shows that this study could use ORME-DQ (Batini et al. [2]) and Otto et al. [19] as the general process framework (GPF). Both studies explicitly **provide identification of business problems within their proces**s and **the process model** that can be used for this study.

The GPF to develop data quality metrics that will be used within this thesis is the methodology developed by Otto et al. [19]. The selection of this methodology is using these considerations:

i. The methodology is specific for identifying business-oriented data quality metrics.
ii. The components in the process model and metamodel are developed on the basis of several other studies:
   a. The structure of the methodology that consists of activities, results, metamodel, role, and techniques

b. Pragmatic definition and representation, such as the definition of quality dimensions, roles, and measurement methods
c. Relationship with business performance

iii. The methodology provides tangible artifacts to use, for example, data quality requirements, questionnaire, and data quality dimensions.

iv. The methodology provides a metamodel for the process. It is possible to develop the components that suit a certain case study, for example, the data quality requirements, data quality dimensions, or performance assessment.

As explained in 1.5, this thesis will conduct theory testing research to answer the second research question and will need to provide the list of data quality metrics for Elsevier's case to answer the third research question.

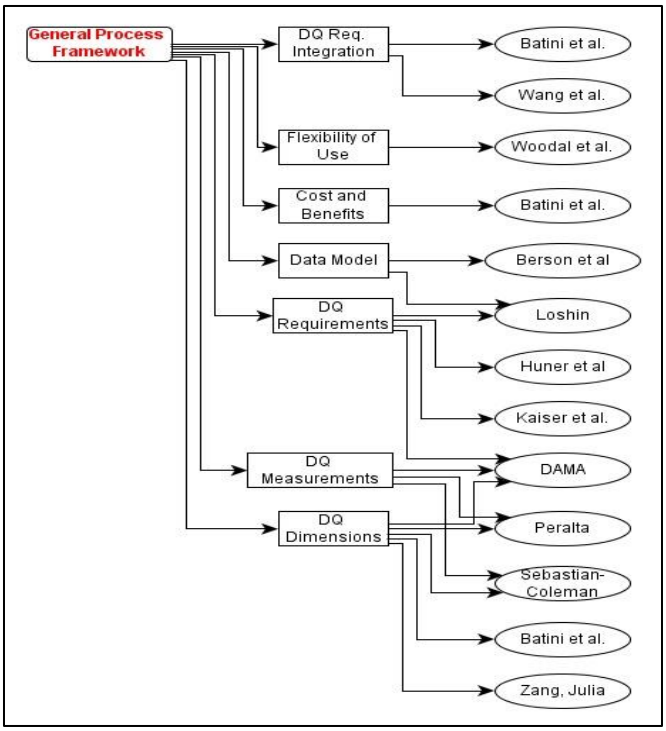


Figure 11 Literature Studies

To make sure that the selected process model is adaptive enough for Elsevier's case, we need to incorporate several other studies that are related to MDM, the components in the process model, and process flexibility. Thus, we still use the results of other studies and correlate them with the selected GPF (Figure 11), as follows:

a. DQ Metrics Requirements (Hüner et al. [12], DAMA [5], and Loshin [15])
   A data quality metrics requirement is the component that is used to develop and test the data quality metrics; that is, the data quality metrics should conform to the requirements. We could develop the list of requirements that meet the need of Elsevier as a publishing company using several studies' results.
b. DQ Dimensions (Zang, Batini et al. [2], Peralta [21], Sebastian-Coleman [23], Morbey [17], and DAMA [5]).
c. DQ Measurements (Pipino et al. [22], Peralta [21], and Sebastian-Coleman [23], and DAMA [5])

DQ dimensions and DQ measurements are two main objects in the assessment and measurement phase (Batini et al. [3]). We need to use the results of several studies to get a complete set because each research usually focuses on several dimensions or measurement methods. It is also possible that they use a different approach or definition for the same dimensions.

d. DQ Methodology Process's Flexibility (Woodal et al. [29])
   This research provided that we could have a variation of process model in the data quality assessment and improvement methodology to adapt for a certain case. It is a useful finding for this study because it is possible that the case study requires some adjustments to the GPF.
e. Data Model in MDM (Berson et al. [4] and Loshin [14])
f. DQ Requirement Integration (Wang et al. [29] and Batini et al. [4])
   The goal of this project is to develop the data quality metrics for the MDM. The data model or requirement integration research provides a solution to ensure the result will be feasible for the MDM system/environment.

3.2 General Process Framework

3.2.1 Goal

The goal of the method is to identify the business-oriented data quality metrics for a product MDM.

3.2.2 Metamodel for the method

The methodology has several important components/entities that need to be identified or developed. The metamodel that covers the required components is as depicted in Figure 12. The activities within the process model have a goal to develop those components. Table 3 provides a more detailed description of the metamodel.

Table 3 Metamodel component description, Otto et al. [19]

No	Component	Description
1	Data	Data is a representation of objects and relationships between objects. The paper considers corporate master data with a focus on values assigned to data elements.
2	Data Defect	Data defect represents the condition where the data does not meet the technical requirement or consumer's need. It is a result of incidents, such as input error, and it poses a risk to data.
3	Data Quality Metrics	A quantitative measure of the degree to which data possesses given quality attributes. In a data quality metric, there are descriptions for related dimension, where to measure and what data to measure, the measurement method, and the scale used for measurement.
4	Business Process/ Process Activity	Sequence of chronologically and typologically linked tasks is intended to generate a clearly defined output, bringing about customer benefit.
5	Process Metrics	A quantitative measure of the degree to which a business process possesses given quality attributes. It provides information about a process's state, indicating weak points and allowing immediate reaction.
6	Business Problem	State or incident leading to decreased business process performance. It poses a risk to a business process and could affect the business goal.

No	Component	Description
7	Preventive Measure	Activities that are conducted to avoid or lower the probability of data defects.
8	Reactive Measure	Activities that are conducted with the data when a defect occurs.

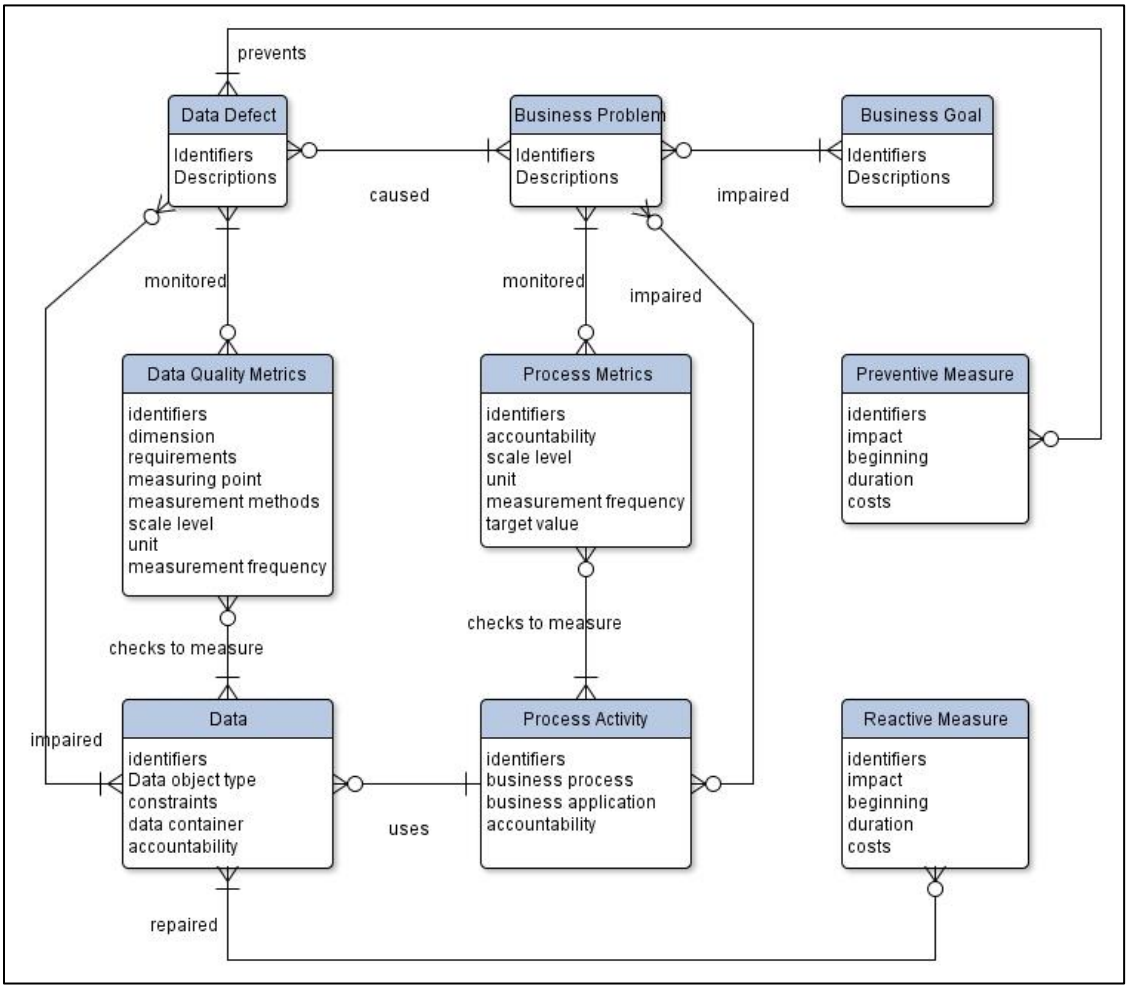


Figure 12 Metamodel for the Method, Otto et al. [19]

3.2.3 Process Model

The process model for this method is as described in section 3.1.1.6 and 3.1.3. The activities that are conducted within each phase will identify or develop the component/entity in the metamodel (Figure 12). The matrix that provides the relationship between the process and the entity is in Table 14. The detailed information about the process is as follows:

i. Phase I

This phase consists of three activities—namely, identify business process and process metrics, identify IT systems, and identify business problems and data defects. The aims of this phase are selecting the business process and metrics to focus and to identify the business experts, identifying the relevant IT systems (e.g., applications, database) and the IT experts, and identifying cause-effect chains between business problems and data defects.

The important result of this phase is an informal documentation of cause-effect chains (i.e., business problems and data defects) and likely affected business processes, process metrics, and data classes.

a. Activity I.1 Identify Business Process and Process Metrics
   This activity aims at identifying business process and process metrics to focus on during the remaining identification process. There should be criteria for the selection of a particular business process; for example, it is important to for the company's business success and the availability of metrics and measurement values. Also, it results a list of contacts that might be interviewed for activity I.3.
b. Activity I.2 Identify IT Systems
   This activity aims at identifying IT systems (e.g., ERP systems, CRM systems, or databases) that are supporting the identified business processes. It also results a list of IT experts that might be interviewed for activity I.3.
c. Activity I.3 Identify Business Problems and Data Defects
   It is the main activity of phase I, and it aims at identifying cause-effect chains between business problems and data defects. There are two methods to identify the cause-effect chains: (i) Identifying causing data defects from identified critical business problems and (ii) identifying potential business problems for already-known data defects. Otto et al. [19] provided interview guidelines and exemplary cause-effect chains to support this activity.

ii. Phase II

This phase consists of two activities—namely, defines and ranks requirements for data quality metrics and specifies data quality metrics. The aims of this phase are to select requirements for the DQ metrics, which consist of generic and company specific requirement, and metric specification (data item, measurement method, measurement point, and measurement scale).

a. Activity II.1 Define and Rank Requirements
   This activity aims to define the requirements for data quality metric specification. It will be used as a guide to define data quality metrics in activity II.2, for example, the scale to be used or the selection of method. The list of requirements will also be used for verification activity (III.1). The list of requirements should comprise both generic (e.g., a specified measurement device and measurement points) and company specific requirements (e.g., facility to visualize metric measurements in a specific manner).
b. Activity II.2 Specify Data Quality Metrics
   It aims to at least specify one data quality metric. This activity comprises the specification of a (subset of a) data class that is measured by the metric, the specification of a measurement device and a measurement point where the measurement takes place, the specification of the measurement scale, the specification of measurement procedures and its frequency. Those sub activities provide information needed for a data quality metric. It is part of data quality metric requirements that have been defined in this study.

iii. Phase III

This phase exists to verify whether the DQ metrics meet the requirements in phase II.

a. Activity III.1 Verify Requirement Fulfillment

This activity verifies the requirements defined in activity II.1. If a requirement is not met, the process starts again with II.1 in order to check the requirements' content and ranking.

b. Activity III.2 Document Data Quality Metrics Specification
The result of this activity is a documentation of the specification of the DQ metrics (activity II.2), including the identified cause-effect chains (activity I.3), and the requirements (activity II.1). This documentation might be used as a requirements document for the implementation of the DQ metrics.

iv. Data Quality Metrics Integration Phase

This phase is required to integrate the data quality metrics specification from several applications into a product MDM system. Wang et al. [29] provided the process in for data quality metrics integration as described in section 3.1.3. Because the result of step 1-3 is similar with the Otto et al. [19], this study focuses on the fourth step, namely conduct quality view integration. The fourth step in the process model is using the process model for schema integration by Batini et al. [4] as described in Figure 13.

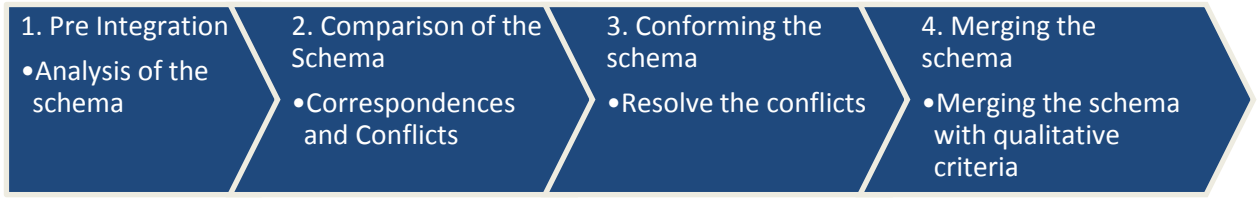


Figure 13 Schema Integration, Batini et al. [4]

The activities for the schema integration process are as follows:

1. Pre-integration
   An analysis of schemas is carried out to provide the information needed to make these decisions: the number of schemas to be integrated, the amount of designer interaction, the order of integration, and a possible assignment of preferences to entire schemas or portions of schemas.
2. Comparison of the Schema
   The schemas are compared to determine the correspondence among concepts and possible conflicts.
3. Conforming the Schema
   The goal of this activity is to resolve the schema conflicts with the designers and users before merging them.
4. Merging and Structuring
   The consideration for merging activity is using several qualitative criteria, namely completeness and correctness, minimality, and understandability.

4 Empirical Evaluation in Elsevier

As described in section 1.5, the step after selecting the solution process is to develop the solution where the goal is to identify the data quality metrics that have business impact using the process model specified in section 3.2. The thesis work conducts the theory testing research where the aim is to test and make adjustments (Verschuren et al. [27]) if necessary to the general process framework (GPF). The aim of this section is to answer the second and third research questions:

- *How appropriate is the methodology for a practical case? What processes or components should be altered?*
- *What are the data quality metrics specifications for a study case in Elsevier?*

The details of the activities and results could be found in Appendix 7, Appendix 12, and Appendix 13. Combining the GPF for developing data quality metrics by Otto et al. [19] and data quality integration by Wang et al. [29], we could develop a process to develop data quality metrics for the MDM, as in Figure 14. The list of data quality metrics is developed for each process or application, and the results will be integrated to get the data quality metrics for the product MDM repository.

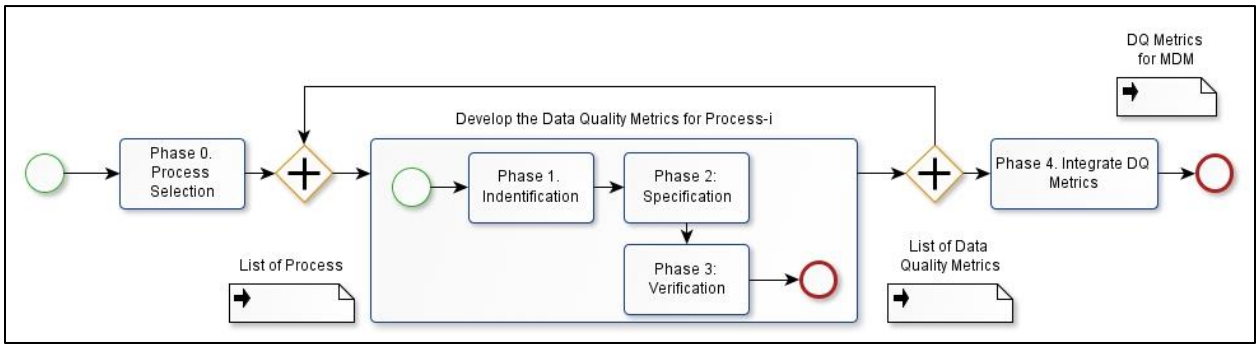


Figure 14 DQ Metrics Development Process for MDM

4.1 Phase 0. Process Selection

This process is part of the business process and process metrics activity of "Phase I. Identification" in the GPF. This process is needed in the Elsevier case because the documentation of the business process is incomplete and the knowledge of each process is dispersed. However, this process selection activity is also useful for MDM implementation with an iterative model.

The process model in GPF describes that the selected process should have metrics and measurement values to enable a comparison between poor data and process's performance [19]. This activity uses several other considerations to select the case study as follows:

i. It should be a data consumer process to provide useful and usable information (Lee et al. [13]).
ii. It is an important business process on the basis of company strategy or revenue.
iii. The complexity of the process can be reduced for research purpose.
iv. The data model is mature.

Thus, the selected processes are for books and journals e-commerce (e-store) that cover data production and setup, and marketing activity as in Figure 15.

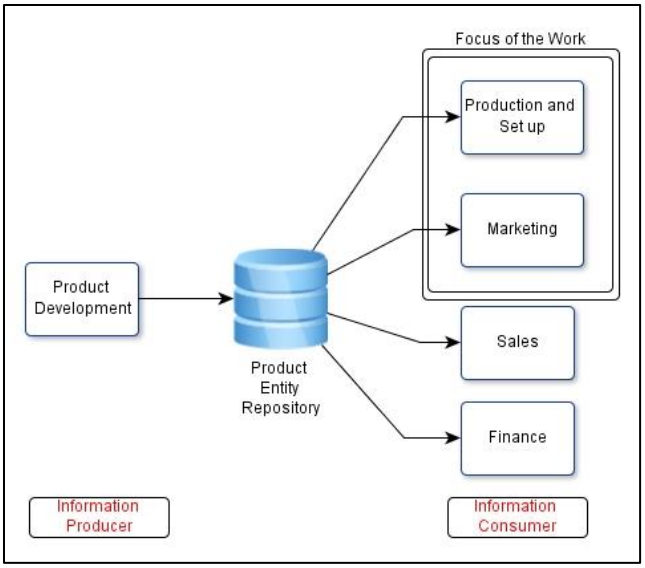


Figure 15 Business Processes and Product Entity Repository

4.2 Phase I. Identification

The activities in this phase are mainly literature studies, interviews, and workshops. Because the main goal is to identify the business problems and the causing poor data, we need to use the similar definition of data quality dimensions at early phase. This study is using the data quality dimensions defined by Morbey [17] (Table 16) and Zang (Table 18) because they used the practitioner's perspective. Since each data quality study only focuses on several dimensions or measurements, it is important that we also map the data quality dimensions with the ones developed by other researchers, as in Table 4.

Table 4 DQ Dimension Mapping

No	Dimensions	Batini	Coleman	Peralta	DAMA	Zang
1	Completeness per row (horizontal completeness)	Completeness	Completeness	-	Completeness	Completeness
2	Syntactical correctness (conformity)	-	-	Syntactic Correctness in Accuracy	-	Validity
3	Absence of contradictions (consistency)	Consistency	Consistency, Consistency for Validity	-	Consistency	Integrity, Consistency
4	Accuracy including currency	Accuracy	Accuracy by (in)Validity	Accuracy	Accuracy	Accuracy, Timeliness
5	Absence of repetitions (free of duplicates)	-	-	-	Uniqueness	Duplication
6	Business referential integrity (integrity)	-	Integrity	-	Referential Integrity	Integrity

No	Dimensions	Batini	Coleman	Peralta	DAMA	Zang
7	Completeness (Cross-check sums, vertical completeness)	Consistency	Consistency for Integrity	-	-	Consistency
8	Normative consistency	Consistency	Consistency	-	Consistency	Consistency

4.2.1 Identify Business Process and Process Metrics

This activity has the same goal defined by the GPF—namely, to determine the appropriate process, identify the responsible persons, and identify the process metrics or KPI. Because the main process has been selected in phase 0, we need to identify the sub processes or activities that are mostly affected by the product data quality. This thesis selects the product availability and marketing activities in Figure 16 because the result of those activities could determine whether a product data could be displayed correctly on the e-commerce websites, whether a potential customer could find an appropriate product page, and whether those customers will buy the product. Those activities should use the product data that are usable and useful (Lee et al. [13]), and the data quality should constantly meet the end customer's expectation (English, [7]).

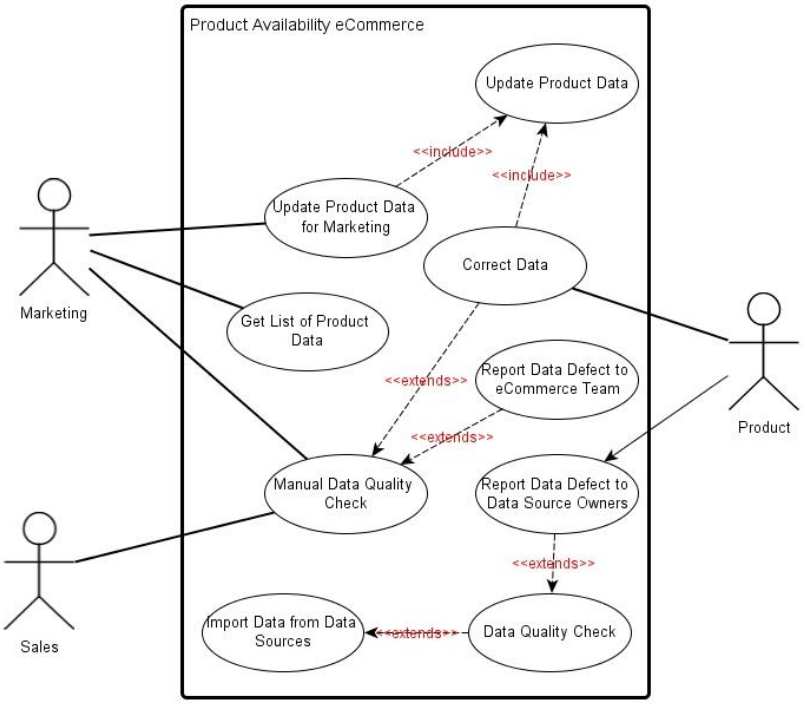


Figure 16 E-commerce Use Case

The GPF also provides the list of roles in the company that should be involved in a certain activity of developing the metrics. This is important because some of the roles are attached to the same individuals, and we need to develop the information using the perspective of the correct role. The client, process owner, and data user are attached to the same individuals even though they could have

different responsibilities and requirements. An example is that the process owner should provide the correct data, while the data user should get the correct data. The absence of a specific role like the technical data steward also makes everyone act as the data steward.

In the GPF's metamodel, we could find the relationship between the business problems and process metrics or KPI. However, the case study does not provide the list of KPIs for processes in e-commerce. We need to develop a new list of KPI using other relevant studies because it is a required component in the metamodel. The KPI in this thesis provides the relationship between the financial and the internal process activities, namely the product availability and the marketing.

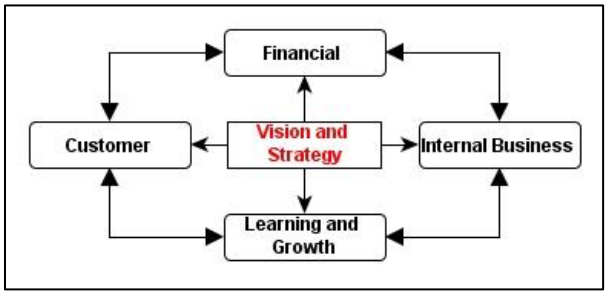


Figure 17 BSC Framework

Tsai et al. [26] developed e-commerce metrics (KPI) using a balanced scorecard framework (BSC, Figure 17) and an empirical study by interviewing the experts. The result of the interview is processed using the Delphi method, and the selected metrics are as described in Table 28. This study selects the metrics for the study case as follows:

a. This study selects only the metrics under the customer and internal business components with these considerations:
   - The BSC by Kaplan in Figure 17 describes that that customer and internal business components have a direct impact on the financial component.
   - Model 1 of Pearlman [20] in Figure 18 shows that components that directly affect the financial component are the customer and the internal business.
   - Tsai et al. [26], using DEMATEL, also provides that the customer is the most impacted component, and internal business is the component that gives the most impact.

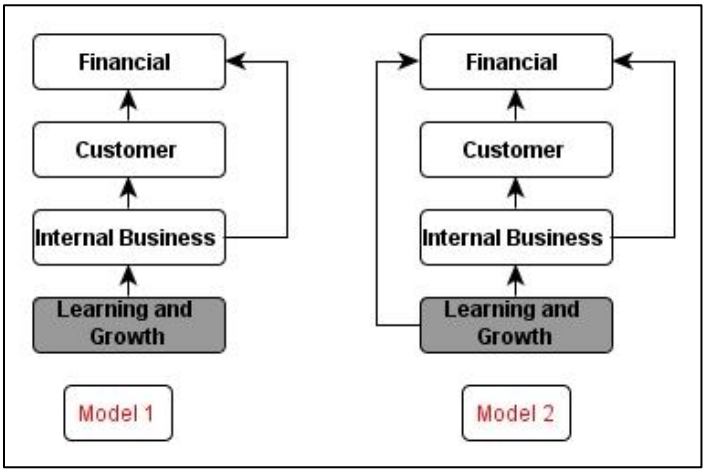


Figure 18 BSC Component Relationship, Perlman [20]

b. The metrics under customer and internal process are not directly affected by product data quality, for example, payment function, rapid delivery, and transaction safety and assurance. This study selects **only the metrics that are related to product data**, as in Table 5.

Table 5 Selected Metrics for E-commerce, Tsai et al. [26]

KPIs	Mean	Median
Customer (C2)		
Willingness to purchase	8.60	9.00
Product information	7.40	7.00
Increase in trust from customers	7.85	7.00
Search engine optimization	7.50	8.00
Internal Process (C3)		
Ability to write marketing proposals	7.55	7.00
Ability to conduct Internet marketing	8.50	9.00
Selection of products for display	8.40	9.00

4.2.2 Identify IT Systems

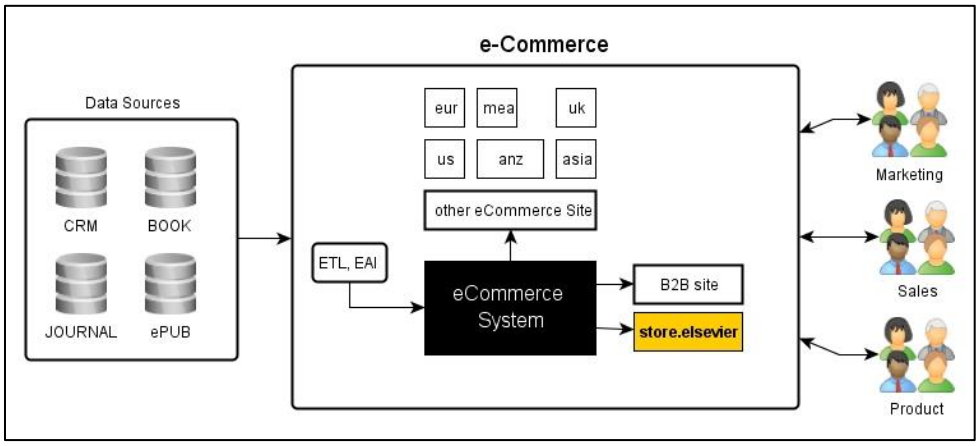


Figure 19 E-commerce System Context Diagram

This activity has the same goal defined by the GPF, namely to identify the business applications and data model. The challenge in this phase is to determine the appropriate data model because each unit could have their own data model like the Web data model, the data model in the data exchange process, and the data model in the databases.

This thesis work uses the Web data model for book (Figure 20) and journal (Figure 21) because they are used by the end customer to decide whether they are on the correct product page and to decide for buying one. Using the comparison of information between the internal e-commerce sites and the competitors in Appendix 7, we could also use these models as a reference for a generic e-commerce process.

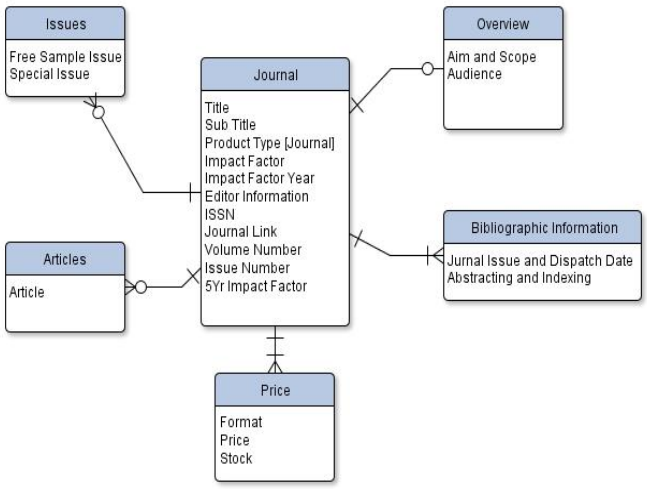


Figure 20 Journal Data Model in E-store Website

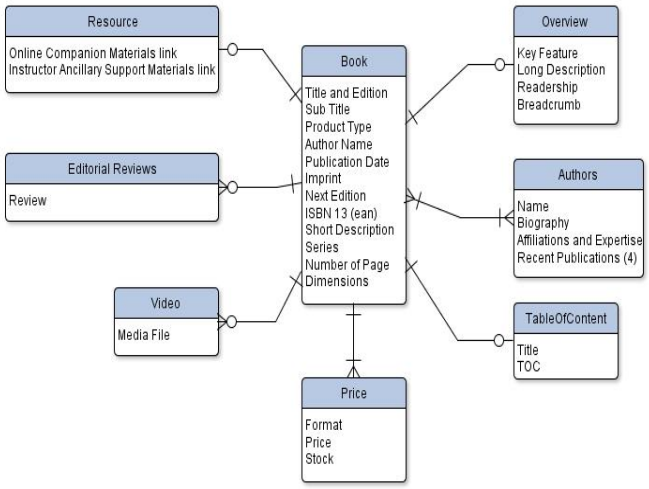


Figure 21 Book Data Model in E-store Website

4.2.3 Identify Business Problems and Data Defects

This study identifies the business problems and data defects in Appendix 4 using the top-down (business problems first) and bottom-up (possible data defects first) approaches, but the top-down approach is more successful. This condition is also found in the GPF description. Hüner et al. [12] mapped the business problems with the internal company's KPI and assessed them with the performance report to assess the business impact. The case study does not have a set of KPI and does not have performance assessment related with poor data. This study uses several approaches to validate the business problems as follows:

a. Validate data quality issues and e-commerce performance using literature studies

Within the Molla and Licker [16] e-commerce system success model, the content quality and system quality are the key factors providing customer satisfaction that leads to a purchase order. Some attributes for the content quality are accuracy, currency, and completeness. Flanagin et al. [7] stated that the up-to-dateness and completeness of information are among the top 3 factors to determine credibility or trust of commercial information. The survey by Clavis Technology[2] in 2013 also showed that incomplete and inaccurate information about e-commerce could lead to a frustrating experience for buyers, low conversion rates, and lost sales opportunities. Thus, we could conclude that the data quality defects developed in the workshop affect the e-commerce performance.

2

http://www.gs1us.org/DesktopModules/Bring2mind/DMX/Download.aspx?Command=Core_Download&EntryId=701&PortalId=0&TabId=785

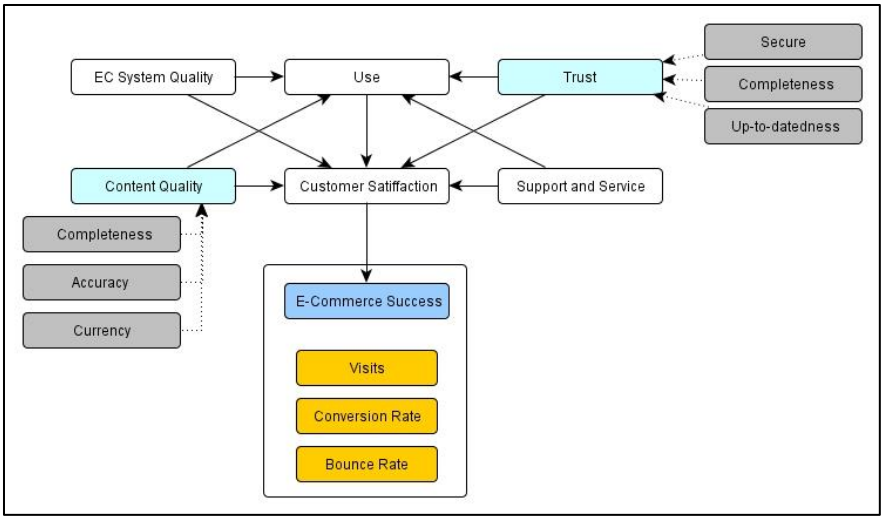


Figure 22 Information Quality and E-commerce, Molla and Licker [16], Flanagin et al. [7], Clavis

b. Filter the business problems by mapping them with the KPI, as in Table 6. It shows that all business problems are mapped with at least a KPI.

Table 6 Business Problems – E-commerce KPI Mapping

KPIs	Business Problems
Customer	
Willingness to purchase	(i), (ii)
Product Information	(i), (iv), (v)
Increase in trust from customers	(i),(v)
Search Engine Optimization	(iv)
Internal Process	
Ability to write marketing proposals	(iii)
Ability to conduct internet marketing	(iii), (vi)
Selection of products for display	(ii), (v)

c. Develop a performance assessment and develop the qualitative reasoning of having the poor data quality mentioned in the business problems.

An assessment using six months (08/2013–01/2014) Google Analytics data is also conducted to support the developed business problems.

Table 7 E-commerce Performance Assessment

Attribute	Description	Result
Acquisition	KPI	1.6m visitors. This exceeds the expectation set in 2012 (1.9m/year).
Conversion	KPI	0.91%. This is **below** the target in 2012 (1.1%) and **below** the retail average, which is 3% in 2012.[3]
Average Order Value	KPI	The average order size is USD 147, larger than expectation in 2012 (USD 96.04).

[3] http://www.marketingsherpa.com/article/chart/average-website-conversion-rates-industry

Attribute	Description	Result
Stickiness	KPI	Bounce rate is 71.8%, **larger** than expected in 2012 (50%).
Search Terms	Additional Information	▪ Within Top 25 search terms in Google search, 15 are imprints and 7 are titles. Only 2 terms are subject. ▪ Search terms that have high (>-20%) click-through rates (CTR) are imprints and titles. ▪ A term with author and title could **lower** the CTR into 9%.

The assessment result in Table 7 shows that the performance of e-store should be enhanced to get a higher conversion rate and a lower bounce rate. There could be several reasons for a visitor to cancel the transaction or leave a certain page without further actions. As explained in Figure 22, the low quality of information could lead to frustrating experience, low conversion rate, and lost sales opportunity. A simple assessment using a data quality tool on the marketing dataset and Google Analytics report in Table 29 informs that poor data quality contributes to e-commerce performance. The result reveals that some data in e-store are incomplete, incorrect, and inaccurate.

4.2.4 Overall Process Model and Metamodel

i. Process Model

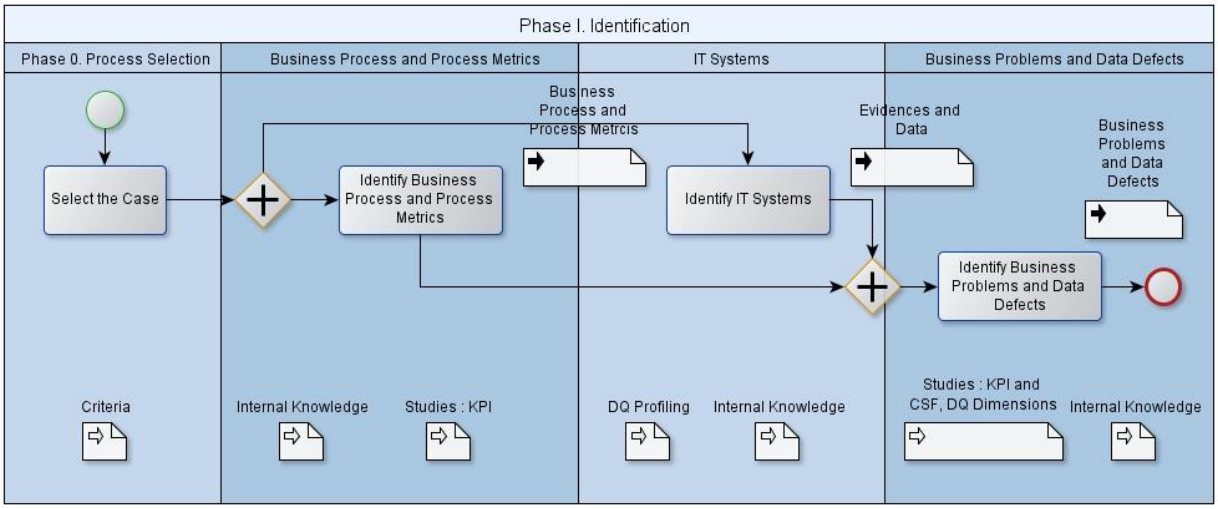


Figure 23 Phase I. Identification Process Model

An alternative process model that can be used on the basis of the solution development process in the study case is as in Figure 23. The description of the alternative process that differs from the GPF is as follows:

a. "Phase 0. Process Selection" to select the cases is required prior to any activities in the process model. The process model in the GPF is conducted for each identified case in phase 0 that meets the criteria.
b. Identify business process and process metrics activity, identify IT systems activity, and business problems and data defects activity are conducted in parallel because they share the same

counterparts in the company and make the process effective in time. Each activity is an iterative one and has several sub activities, namely preliminary interview document to capture the components required in the metamodel, interview document preparation to capture the essential information about the required component, interview process, follow-ups, and validation/confirmation.

c. The activity should allow the use of process metrics or KPI from other sources because it is possible that the business process does not have documented or detailed metrics.
d. The identify IT systems activity should wait for the result of identify business process and process metrics activity to complete for these reasons:
   - To synchronize the information about the business process with the functionalities and data in the IT system
   - To provide a set of DQ profiles for involved attributes as a possible evidence for business problems
e. The business problems and data defects activity should wait for the other two activities to complete for these reasons:
   - To ensure it considers all business processes in the case
   - To get the proper data attributes that represent the data that is a defect. An example is at the first interview there is a business problem caused by the subject of the book. Having identified the IT system, the attribute related to the subject is “category”
   - To get enough evidence (e.g., DQ profiling results) and supporting references (e.g., KPI and CSF from other studies) to retain and validate the business problems
f. The business problems should be filtered and validated. The proposed validation methods are finding supporting literature studies/references, mapping with the KPIs, and developing a performance report.

ii. Components for Metamodel

This phase provides the components defined by the GPF as follows:

a. A list of business processes in e-commerce that is affected by product data quality in Figure 16.
b. A list of KPIs that can be used for e-commerce and related to product data quality in Table 5.
c. The product data model required in the e-commerce system in Figure 20 and Figure 21.
d. A list of data quality dimension definitions that can be used to get the same understanding among interviewees/parties in Table 16 and Table 18.
e. A list of business problems and data defects for e-commerce that is related to product data quality in Table 21. The reactive and preventive measures to maintain the data quality are in Table 22.

4.3 Phase II. Define/Specify

4.3.1 Specify Requirement for DQ Metrics

This study develops and ranks the DQ metrics requirements in Table 8 by combining the requirements by DAMA [5], Heinrich et al. [11], Sebastian-Coleman [23], and Loshin [15]. The rank is decided by the business and information system experts in Elsevier. This method is introduced within the study by Hüner et al. [12], the main reference for the GPF. A list of requirements should be developed for two purposes, namely to guide the development of data quality metrics and to assess

whether the data quality metrics are acceptable. The data quality metrics requirements include the requirements for the values and methods, and whether they answer some business problems.

Table 8 DQ Metrics Requirements

No	Requirement	Code	Description	Reference	Importance	
	Generic					
1	Business Relevance	DQ-R-01	Every data quality metric should demonstrate how meeting its acceptability threshold correlates with business expectations.	DAMA [5]	2	1
2	Controllability	DQ-R-02	The assessment of the data quality metric's value within an undesirable range should trigger some action to improve the data being measured.	DAMA [5]	1	1
3	Acceptability	DQ-R-03	Base the determination of whether the quality of data meets business expectations on specified acceptability thresholds	DAMA [5]	2	1
	Value related					
4	Measurability	DQ-R-04	A data quality metric must be measurable and should be quantifiable within a discrete range	DAMA [5]	2	2
5	Normalization	DQ-R-05	An adequate normalization is necessary to ensure that the values of the metrics are comparable. In this context, DQ metrics are often ratios with a value ranging between 0 (perfectly bad) and 1 (perfectly good)	Heinrich et al. [11]	2	2
6	Interval Scale	DQ-R-06	This means that the difference between two levels of DQ must be meaningful.	Heinrich et al. [11]	2	2
	Methods related					
7	Feasibility	DQ-R-07	It is also required that the measurement procedure can be accomplished at a high level of automation.	Heinrich et al. [11]	1	1
8	Reproducible	DQ-R-08	To produce consistent measurement results and to understand any factors that might introduce variability into the measurement.	Sebastian-Coleman [23]	1	2
	Value and Methods					
9	Aggregation	DQ-R-09	The metrics must allow aggregation of values on a given level to the next higher level.	Heinrich et al. [11]	2	1
10	Comprehensible and Interpretable	DQ-R-10	The DQ metrics have to be comprehensible; for example, considering a metric for timeliness, it could be interpretable as the probability that a given attribute value within the database is still up-to-date	Heinrich et al. [11], Sebastian-Coleman [23]	1	2

4.3.2 Specify Data Quality Metrics

Metric specification is developed by combining several studies' contents of the DQ measurement method. Since each study on data quality could only focus on several attributes of DQ metrics in Table 32 and only for several types (e.g., completeness and correctness), this study combines the results of several studies to develop a complete set of data quality metrics to assess the data quality dimensions found in the business problems. The heuristic that is used by this study is as follows:

i. Phase I, create the initial set of data quality metrics using these activities:
   a. Assess the similarity of data quality dimensions/attributes' definition as found in Table 4.
   b. Assess several possible measurement methods using the result of item (a) as in Table 33.
   c. Assess the attributes using the data quality metrics requirements to make necessary modifications, for example, the normalization and discrete range values.

   The result of this activity is a list of data quality metrics in Appendix 10 with 7 of 11 complete attributes, namely identifiers, dimension, measuring method, scale level, unit, measurement frequency, and requirements. Because the attributes are not attached to a specific application, it can be used for other purposes with the same data quality dimension, such as, completeness per row or syntactical correctness. An example is the DQ metrics for CPR-01 in Table 9.

Table 9 An Example of DQ Metrics in Phase I

No	Attributes	Description
1	Identifier	CPR-01
2	Dimension	Completeness per Row, (in)Accuracy
3	Measurement method	**result** = 1 - (Number of row with empty non null able field divided by the number of all rows)
4	Scale level	Simple ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best = 1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08

ii. Phase II, create a set of refined data quality metrics using these activities:
   a. For each business problem and preventive/reactive measure, assess the appropriate metrics as in Figure 24.

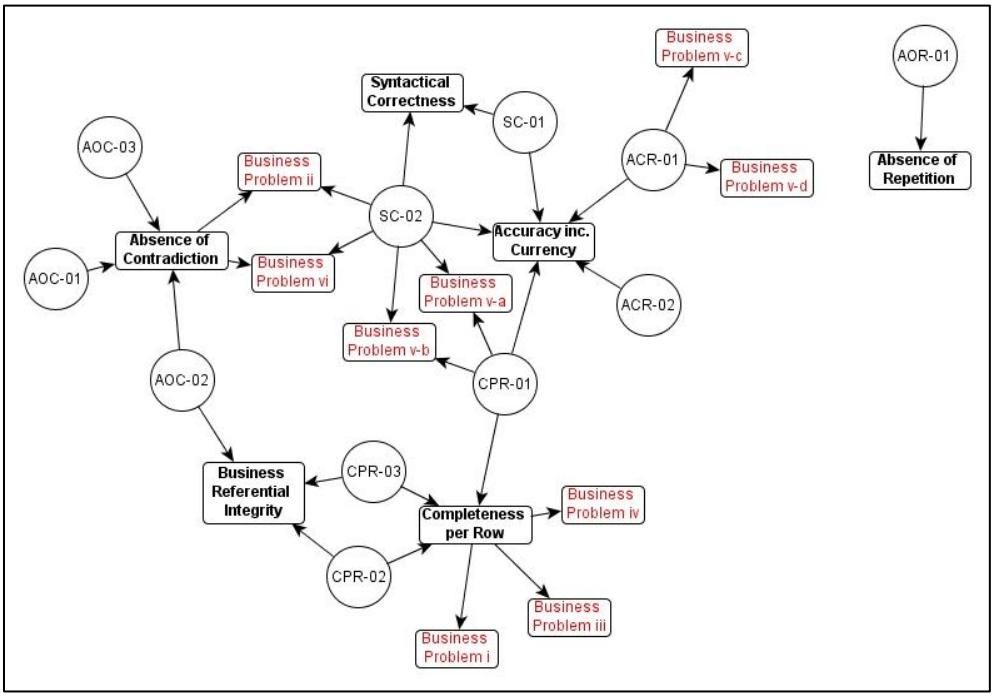


Figure 24 Metrics Relationships

b. Complete the measuring point, data, and data defect attributes of each metric.

As described in the study by Hüner et al. [12], we need to assess the data using the developed DQ metrics to determine the threshold. The data quality metrics specifications are categorized into two types in this study with regard to the sources of the requirements, namely the business problems and data defects (Table 34), and the preventive and reactive measures in the e-commerce system (Table 35). An example of the developed DQ metrics in this phase is CPR-01 in Table 10.

Table 10 An example of DQ Metrics in Phase II

No	Attributes	Description
1	Identifier	CPR-01
2	Dimension	Completeness per Row, (in)Accuracy
3	Measurement method	**result** = 1 - (Number of row with empty non null able field divided with the number of all row)
4	Scale level	Simple ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best = 1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
8	Measuring point	E-commerce database
9	Data	All attributes in Web data model
10	Data defect	Incomplete information in Web data

4.3.3 Overall Process Model and Metamodel

i. Process model

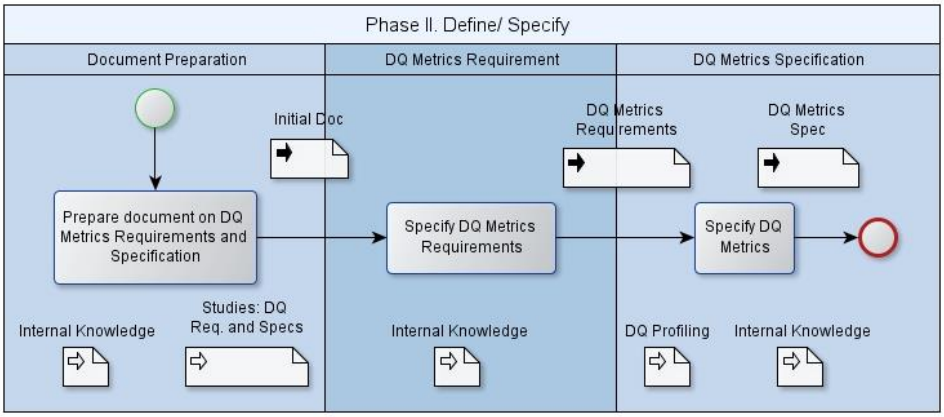


Figure 25 Phase II. Define/Specify Process Model

An alternative process model that can be used on the basis of the solution development process in Elsevier's case is as in Figure 25. The description of the alternative process that differs from the GPF is as follows:

a. Each activity in phase II has these sub activities: preliminary discussion document to develop the components required in the metamodel, discussion document preparation to develop the essential information about the required component, discussion process, follow-ups, and validation/confirmation

b. The document preparation activity is important to make the overall process effective and efficient. Using the requirements and specifications that other studies have resulted provides more credibility of the content. The workshop activities to specify the requirements and specifications only make minor changes.
c. It is important to map the data quality dimensions/attributes' definition from several studies. The reason is because each study usually focuses on several dimensions or measurement methods. The combination of those studies' results could provide a complete set of required data quality attributes.

ii. Components for Metamodel

This phase provides the components defined by the GPF as follows:

a. A list of data quality requirements in Table 8 that is relevant for product data quality in e-commerce.
b. A list of data quality metrics specifications in Appendix 11that is relevant for product data quality in e-commerce case. Each metric has complete information for these attributes: identifiers, dimension, measuring method, scale level, unit, measurement frequency, requirements, measuring point, data, and data defect.

4.4 Phase III. Verify

The DQ metrics developed in phase II should be verified to determine how it meets the requirements in Table 8. The fulfillment to several requirements could be assessed directly by analyzing the metrics development process or the metrics' attributes, for example, the business relevance and the normalization features, whereas some others need a data analysis activity for assessment, for example, the feasibility and acceptability features.

4.4.1 Develop Criteria for Requirements

The criteria for each requirement need to be developed to quantify the level of requirement fulfillment of each DQ metric. The criteria could also be used to determine the evaluation methods for each requirement. The criteria developed in Table 11 have the same value range (1, 2, or 3), and they define the evaluation method, for example, assessment of the metric attributes and assessment against e-commerce database.

Table 11 DQ Metrics Requirements Criteria

No	Requirement	Code	Valuation Criteria	Method
	Generic			
1	Business Relevance	DQ-R-01	3 = Has a related business problem. 2 = Has related preventive/reactive measures. 1 = Has no related business problem or preventive/reactive measures.	Assess the DQ Metrics value [dimension, data, data defect]
2	Controllability	DQ-R-02	3 = There is an action to conduct for a certain DQ value. 1 = The possible action to conduct for any DQ value is unavailable.	Assess the DQ Metrics value [dimension, data, measuring point, method]

No	Requirement	Code	Valuation Criteria	Method
3	Acceptability	DQ-R-03	3 = The threshold value is derived from a performance report assessment. 2 = The threshold value is derived from best practices.	Performance assessment to assess the correlation
	Value related			
4	Measurability	DQ-R-04	It is part of 5 and 6.	
5	Normalization	DQ-R-05	3 = Normalized value in 0–1 2 = Not normalized value.	Assess the DQ Metrics value [unit]
6	Interval Scale	DQ-R-06	3 = Ratio scale. 2 = Interval scale. 1 = Ordinal scale.	Assess the DQ Metrics value [scale level]
	Methods related			
7	Feasibility	DQ-R-07	3 = Could be developed using simple tasks like SQL query. 2 = Need to use a simple programming (e.g., PL/SQL) to develop. 1 = Cannot be developed.	Database Assessment
8	Reproducible	DQ-R-08	3 = Tested. 1 = Not tested.	Database Assessment
	Value and Methods			
9	Aggregation	DQ-R-09	3 = Aggregation at row, table, and database level. 2 = Aggregation at field, table, and database level. 1 = No aggregation.	Database Assessment
10	Comprehensible and Interpretable	DQ-R-10	It is an aggregate of other requirements.	

4.4.2 Verify Requirement Fulfillment

This activity is conducted on the basis of the criteria in Table 11. The requirements for business relevance (DQ-R-01), controllability (DQ-R-02), normalization (DQ-R-05), and interval scale (DQ-R-06) could be assessed by analyzing the value of data quality metrics. An example is whether a specific metric answers the business problems or preventive/reactive measures (Figure 24). While the requirements for acceptability (DQ-R-03), feasibility (DQ-R-07), reproducibility (DQ-R-08), and aggregation (DQ-R-09) should be assessed using a database assessment activity.

i. Database Assessment

The DQ metrics assessment against the database is needed to evaluate these requirements: acceptability (DQ-R-03), feasibility (DQ-R-07), reproducibility (DQ-R-08), and aggregation (DQ-R-09). It is also useful to determine the threshold values for several metrics. The threshold values could be defined manually or automatically as follows:

- Manually
  a. Using a value on the basis of the importance of the field and DQ attributes in the application. This is for accuracy-related measurements.
  b. Assessment result against the sales and Google Analytics data.
- Automatically
  Use assessment results that are considered historical data. An example is using the mean and standard deviation value from three or more data quality assessment activities (Sebastian-Coleman [23]).

The result of the database assessment using the developed DQ metrics is in Table 12 (complete result is in Appendix 12). All of the DQ metrics are tested except for AOC-03, a metric part of absence of contradiction, because the data is not available.

Table 12 DQ Metrics Assessment Result Summary

No	DQ Attributes	Assessment Result		Threshold 1	Threshold 2	Description for Threshold
		Book	Journal			
	Business Problems					
1.	Completeness	0.9265	0.8364	0.8	0.9	On the basis of GA and sales assessment.
2.	Consistency (Title-Category)	0.7908	1.00	0.8	0.9	Using completeness.
3.	Accuracy (Location-Price)	0.9989	1.00	1	1	▪ It must be accurate 100%. Inaccurate data will result to non-displayed or non-fulfilled products. ▪ Use historical data: mean + standard deviation of 3 assessment results.
4.	Accuracy (Location/Format Type - Fulfillment Company Code)	0.9552	0.9261	1	1	
5.	Accuracy (ISN)	-	0.8750	1	1	
	Preventive and Reactive					
6.	Completeness per Row	0.9265	0.8364	0.8	0.9	See 1.
7.	Syntactical Correctness	0.9754	0.9989	0.9754	0.9754	Use historical data: mean + standard deviation of 3 assessment results.
8.	Absence of Contradiction	0.9542	0.9824	0.9340	0.9340	
9.	Absence of Repetition	1.00	1.00	1	1	
10.	Accuracy including Currency		0.9139	0.9139	0.9139	

ii. Evaluation of DQ Metrics and Validation

The evaluation of each DQ metric is conducted using the criteria for each requirement as in Table 13. An example assessment for CPR-01 is as follows:

- Business relevance is 3 because it causes several business problems.
- Controllability is 3 because we could add some information if the completeness is low.
- Acceptability is 3 because the threshold value is derived from a performance report assessment.

- Normalization is 3 because the assessment result is always between 0 and 1.
- Interval scale is 3 because the assessment result is in ratio scale.
- Feasibility is 2 because we need to create a simple application to assess; that is, it cannot be assessed using only query command.
- Reproducibility is 3 because it has been tested using a database.
- Aggregation is 3 because the measurement is aggregating at row, table, and database level.

Table 13 DQ Metrics Assessment Result

No	Requirement	Value	CPR-01	CPR-02	CPR-03	SC-01	SC-02	AOC-01	AOC-02	AOC-03	ACR-01	ACR-02	AOR-01
	Generic												
1	Business Relevance	3.5	3	3	3	2	3	3	3	2	3	2	2
2	Controllability	4	3	3	3	3	3	3	3	3	3	3	3
3	Acceptability	3.5	3	3	3	2	2	2	2	2	2	2	2
	Value related												
5	Normalization	3	3	3	3	3	3	3	3	3	3	3	3
6	Interval Scale	3	3	3	3	3	3	3	3	3	3	3	3
	Methods related												
7	Feasibility	4	2	3	3	2	3	2	2	2	3	3	3
8	Reproducible	3.5	3	3	3	3	3	3	3	1	3	3	3
	Value and Methods												
9	Aggregation	3.5	3	2	2	2	2	2	2	2	2	2	2
TOTAL		28	2.85	2.87	2.87	2.48	2.75	2.60	2.60	2.23	2.75	2.62	2.62
% Score			95.24	95.83	95.83	82.74	91.67	86.90	86.90	74.40	91.67	87.50	87.50

*Requirements number 4 and 10 are not assessed because each is an aggregate of other requirements.

Furthermore, we could use mean and standard deviation function to select the metrics with acceptable value. On the basis of the chart in Figure 26, only AOC-03 has the value below the mean and standard deviation value. The reason for the low value for the metric is it is not tested because of data unavailability.

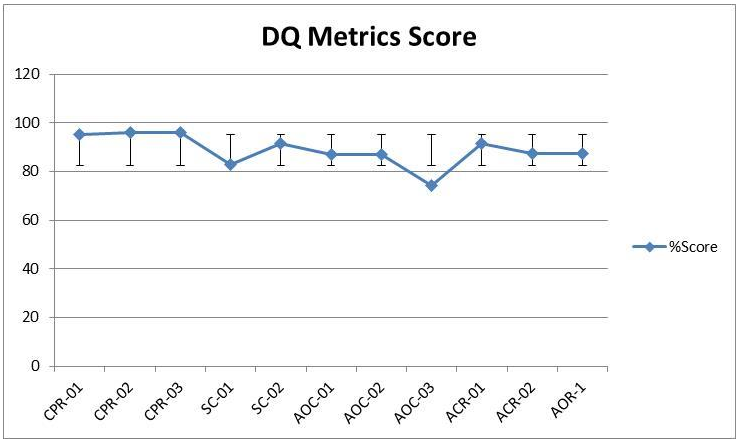


Figure 26 DQ Metrics Score

4.4.3 Overall Process Model and Metamodel

i. Process Model

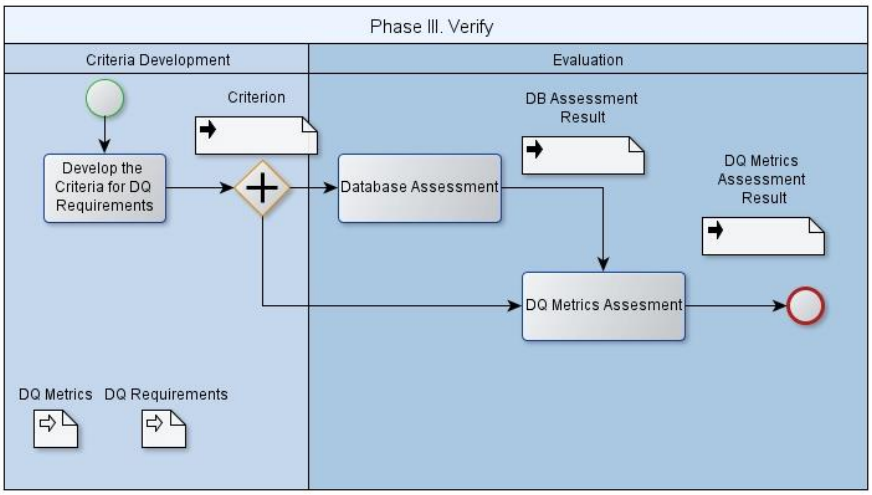


Figure 27 An alternative for Phase III. Verify Process Model

An alternative process model that can be used on the basis of the solution development process in Elsevier's case is as in Figure 27. The description of the alternative process that differs from the GPF is as follows:

a. The verification activity in phase III has these sub activities: preliminary discussion document to develop the components required in the metamodel, discussion document preparation to develop the essential information about the required component, discussion process, follow-ups, and validation/confirmation.
b. The criteria for each requirement should be developed before any assessment. The criteria will provide the activities needed for assessment like a DQ metrics attribute analysis or a database data quality assessment. To quantify the assessment, we also need to have a set of common values for each requirement (e.g., 1–3) where each value is attached to a single criterion.
c. The database DQ metrics assessment is required since some requirements need this activity to assess the DQ metrics fitness level, for example, the feasibility and reproducibility.
d. The DQ metrics are evaluated for their fulfillment to each criterion of the requirements. A further selection of acceptable DQ metrics could be made by having a simple statistic method like mean and standard deviation.

ii. Components for Metamodel

This phase provides the components defined by the GPF as follows:

a. A list of data quality requirements criteria and values in Table 11 relevant for product data quality in e-commerce.
b. A set of acceptable threshold values for DQ metrics in Table 12 for product data quality in e-commerce.
c. A list of acceptable data quality metrics specifications in Table 13 for product data quality in e-commerce.

4.5 Phase IV. Data Quality Metrics Integration

As described in 1.5, this thesis work has two cases—namely, e-commerce system that is mainly used in the validation phase and the Elsevier customer system. The role of the second case is to develop and test the process model to integrate the data quality metrics for the product MDM. This thesis work also develops the data quality metrics for the second phase where the business problems and data defects are inferred from its data rules. Because Elsevier has developed its product MDM data model, this thesis conducts the integration in two phases (Appendix 13)—namely, the integration of e-commerce system and customer system and the integration of integrated applications with the product MDM data model.

4.5.1 Pre-integration

Some of the results of the GPF's activity are data structure and data quality metrics from the e-commerce system that we need to convert into the data model in Figure 7. This data model, with the cell level tagging, is argued by Wang et al. [29] to meet the needs of multidimensionality and hierarchicality of data quality. They put the tagging at the cell level because the attribute value of a cell is the basic unit of manipulation, and each attribute in the same record could be manipulated at a different point of time from different sources. This definition matches the requirement of the MDM.

4.5.2 Comparison of the Schemas

There are two main components in the data mode, the data structure and quality metrics. This activity also compares the two components to find the possible correspondences and conflicts. Both correspondences and conflicts are found in the mapped attributes like book title and book name. While the unmapped attributes provide a list of conflicts, for example, the impact factor attribute is only found in e-commerce and we need to decide which attributes to add in the product MDM data model.

4.5.3 Conforming and Merging the Schemas

Developing the resolution for the conflicts in this activity is using several qualitative criteria, namely completeness and correctness, minimality, and understandability. We also need to merge the structure for the data model, which consists of column name, data type, and data size, and to merge the other attributes of data quality metrics. The merge metrics attributes are the data and the rules in measurement method.

4.5.4 Overall Process Model and Metamodel

i. Process Model

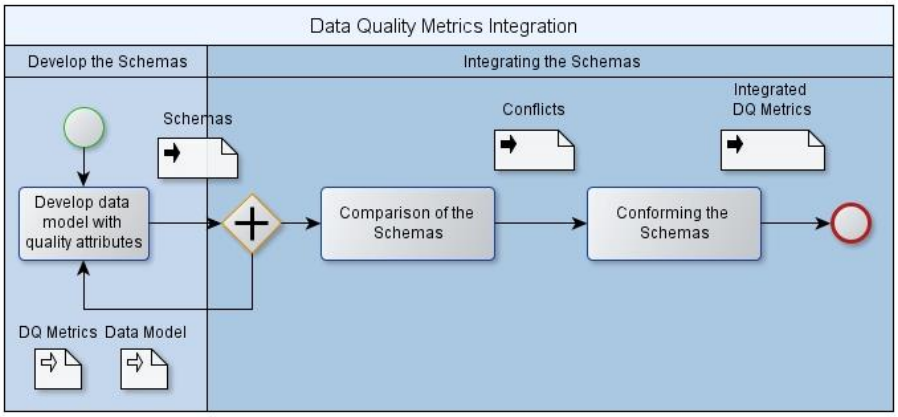


Figure 28 DQ Metrics Integration Process

The data quality metrics integration is conducted using the fourth step of the process model by Wang et al. [29] (Figure 8) and the schema integration model by Batini et al. [4] to develop a list of data quality metrics in Appendix 13that is relevant for product data MDM.

The activities within this process are as follows (Figure 28):

a. Development of the data model with quality attributes
   For each data quality metrics, develop the data model that conforms to Figure 7. Thus, each attribute will have this information: data type, data size, data quality dimension, and data rule.
b. Comparison of the schema
   An assessment for each attribute is conducted to find possible conflicts such as different column name, different data type, and different data rules.
c. Conforming of the schema
   There are common schema conflicts within the two phases, and they are resolved with the qualitative criteria developed by Batini et al. [4], namely completeness and correctness, minimality, and understandability. The conflict resolution is also developed on the basis of two assumptions, as follows:
   [1] The architectural style of the MDM is a transaction hub where master data object updates are in the MDM repository.
   [2] The product data model for the MDM is developed on the basis of the requirement of n (>2) other applications.

The conflicts in conforming the schema and the resolutions are as follows:

a. Different Column Name
   Using the completeness and minimality criteria, we need to select one column name that is considered correct. To determine correctness, we use the understandability criterion. The column name is qualitatively most understood by most applications in the MDM. When comparing the column name, this thesis work selects the one in the product master data model because it has been agreed upon by other n (>2) applications in the company.
b. Different Data Type
   This thesis uses minimality and correctness criteria to resolve the conflict. It selects the most basic form of the data type with the assumption that the value compound process could be conducted at the application level or data exchange.
c. Different Data Size
   This thesis uses the correctness criteria to resolve the conflict by selecting the larger size to avoid data pruning. The impact of this selection is the application adjustment for the ones with smaller data size.
d. Different Data Rules
   This thesis uses the completeness criteria to resolve the conflict because the rules themselves are not conflicting. The data rules are used in completeness, syntactical correctness, and absence of contradiction data quality metrics. There is one rule that is not applicable in the product master data model because the related data attribute is considered not part of the data model in MDM.

e. Non-feasible Measurement Methods

This thesis uses correctness and understandability criteria to determine whether a measurement method in a data quality metric is feasible. The only non-feasible measurement method is ACR-01 for accuracy because it compares the value of an attribute with the ones in the data source. It is not applicable for product MDM because it is the data source in the transaction hub architectural style.

iii. Components for Metamodel

This phase provides a component defined by the GPF—namely, a list of integrated data quality metrics specifications for product data model in MDM in Appendix 13.

5 Conclusion

The main research goal of this study is as follows:

To identify, collect, analyze, and evaluate the quality metrics for a product master data; to allow quantifying and improve their value. This study manages to introduce a process model, heavily based on Otto et al. [19] and Wang et al. [29], to address the main goal. This study also provides a list of data quality metrics for the e-commerce and product MDM system that has business impacts, feasible, reproducible, and acceptable. There are several additional important findings while validating the process in the case study. The findings include the factors in the case study that contributes for alternate process model, the critical success factors in each phase of the process model, the quality of the process model, and the lessons related to some data quality process and product MDM data model in Elsevier.

5.1 Lessons

5.1.1 Contributing Factors for Alternate Process Model

Otto et al. [19] did not provide more detailed information about the case studies other than the data and process domain, thus the contributing factors for the (developed) alternate process model are only for generic features. On the bases of evaluation process in the previous section, the variation in the developed process model could be caused by these features:

i. Data and Process Domain

This thesis uses product and journal data in the e-commerce process, while Otto et al. [19] developed the GPF using three different domains—namely, customer data in customer service process, design data in manufacturing process, and material data in maintenance process. This process is used to develop the data quality metrics for a journal and book product MDM repository and in the e-commerce domain.

ii. Case Study Condition

Otto et al. [19] provided the required information to develop the components in metamodel and provided the list of roles that should be involved in the process model. This thesis managed to develop the required components with these conditions:

- Incomplete documentation for IT, process metrics, and performance reports; thus, we need to develop one using literature studies.
- Several roles are found within a group of people, and each person holds some parts of information. There are also roles that are not formally attached; thus, they attach to almost all personnel like the data stewards.

These conditions require several treatments, for example, the activities should be performed iteratively or in a different sequence, some activities need additional literature studies, or there are some assumptions needed regarding incomplete information.

5.1.2 Critical Success Factors in the Process Model

The main goal of the process model is to develop business-oriented data quality metrics, and each phase of the process has several outputs as its goal. On the basis of the validation process using the case study, we identify the critical success factors for each phase to address its goals.

i. Phase 0. Process Selection

The critical success factor in this phase is the development of criteria to determine the important business processes. The criteria support the function of the MDM, which is to maintain the critical business objects (Loshin [14]).

ii. Phase 1. Identification

The goal of this phase is to discover the business problems and the causing data defects. Using the metamodel, the business problems should affect the business performance. Thus, there are several critical success factors in this phase as follows:

a. The use of the same DQ dimension definitions at early phase to get a common understanding.
b. Identification/development of KPIs.
c. Business problems validation using several activities—namely, mapping with the business KPI, study literatures to support poor data and process performance relations, and performance assessment.

iii. Phase 2. Define/ Specify

The goal of this phase is to develop the data quality metrics requirements and data quality metrics specifications. The critical success factors for this phase are as follows:

a. Using available best practices or literature studies to develop the requirements and specifications. The DQ dimension definition in phase I is also useful to combine several studies for developing a complete set of data quality metrics.
b. The development of DQ metrics requirements is important as a guide to develop the DQ metrics specification and to assess them in the verification phase.

iv. Phase 3. Verify

The goal of this phase is to select the data quality metrics that meet the requirements in phase II. The critical success factors for this phase are as follows:

a. The development of criteria for each requirement is important as a guide to assess the DQ metrics specification.
b. Database assessment is also required to asses several requirements especially for their feasibility and reproducibility. Combined with the performance assessment, database assessment is also useful to determine whether to set the threshold value manually or automatically at first data quality assessment.

v. Phase 4. Data Quality Metrics Integration

The goal of this phase is to develop the data quality metrics in MDM from several application views. The critical success factor for this phase is developing the criteria and rules for conflict resolutions.

5.1.3 Quality of the Process Model

The quality of the process model is assessed using several criteria by Woodall [30] as follows:

i. The practical utility of the approach

The result of the validation process has provided Elsevier with a set of data quality metrics for an e-commerce system and product MDM. The data quality metrics meet the requirements specified by the users, including implementation feasibility, reproducibility, and impact for the business

performance. Thus, we could conclude the alternate process model is feasible to implement for practical cases.

ii. The validity and completeness of the list of activities

The validation process conducts all the activities in the developed process model, which is based heavily on the model by Otto et al. [19] and Wang et al. [29]. It also produces all the components defined by Otto et al. [19]. It means that the activities and the components that are found within the study by Otto et al. [19] and Wang et al. [29] are valid and complete.

The new components/activities are also valid and complete because we could find their implementation within other studies as follows:

a. Criteria for DQ Metrics Requirements

The criteria for the requirements represent the measurement method, scale, unit, threshold, and data of a metric. These attributes are used in the DQ metrics in this study, Otto et al. [19], and Sebastian-Coleman [23]. The purpose of those attributes is to measure an object.

b. Database Assessment

Database assessment is a method of objective measurement using quantitative metrics (Batini et al. [3]). It is also found in several methodologies, for example, DQ-P (Morbey [17]), ORME-DQ (Batini et al. [2]), and hybrid approach (Woodall [30]).

c. Criteria for Conflict Resolution

The criteria for conflict resolution are used in the study by Batini et al. [4] for schema integration.

Because the activities and components in the alternate model are used or validated in this case study and other studies, we could conclude that the developed/alternate process model is valid and complete.

iii. Future resilience of the approach

The process model incorporates the methodology developed by Otto et al. [19] in 2009. They developed the methodology by combining several studies that were developed between 2001 and 2009, including DAMA in 2009, Caballero et al. in 2007, IBM (Alur et al. in 2007), Batini et al. in 2006, and Loshin in 2010.

There is a four years' difference between this study and Otto et al. [19], but the alternate model introduces only small changes of the process, for example, moving the process selection from identification, explicitly defining the KPI development, and explicitly defining the database assessment in verification phase. Thus, we could expect that the developed process is also resilient. However, it is expected for the practitioners to develop new objects for the components in metamodel using new studies' results, for example, DQ dimension by Morbey [17] in 2013 and DQ metrics by Sebastian-Coleman [23] in 2013.

5.1.4 Data Quality Process and Master Data Identification in Elsevier

This study for developing data quality metrics is closely related with master data management and the MDM project in Elsevier. It is expected to work with other components in the company while identifying the data quality metrics in this case study because there are several concerns in MDM, for example, stakeholders' involvement, functional services, data governance, data modeling, data

consolidation and integration, and master data identification (Loshin [14]). This study also provides several additional lessons that can be implemented in Elsevier as follows:

i. Data Quality Process

The two case studies show that the company recognizes the importance of data quality by having data quality checks for several dimensions at data import activity (in-line measurement), for example, completeness and syntactical correctness. However, the business problems show that they also have problems in data accuracy and data consistency that impact the business performance. According to Sebastian-Coleman [23], Elsevier also needs to have a periodical assessment to maintain the data quality for all the records in the repository.

Elsevier should also develop a performance assessment report to assess how the poor data impact their performance and to assess the data quality metrics. They could use the performance report in this study for that purpose.

ii. Master Data Identification

This study also provides a possibility to update the data model in product information management (PIM) with several attributes, for example, table of content and impact factor. Master data are those entities, relationships, and attributes that are critical for an enterprise and foundational to key business processes and application systems (Berson, 2010). Thus, Elsevier should use these criteria to add the attributes in the PIM data model:

a. It is referenced in multiple business areas and business processes (Loshin [14]).
b. It is referenced in transaction and analytical system records (Loshin, [14]).
c. It tends to be static in comparison with transaction systems and does not change as frequently (Loshin [14]).
d. It has low volume volatility (Otto, 2010).

Furthermore, to determine the importance of an attribute, this study suggests using these methods:

a. Information comparison between e-store site and other prominent e-commerce sites for journal/book, for example, Barnes & Noble, Amazon, SAGE, Wiley, and Springer (Appendix 7).
b. Configure Google Analytics to log the tab-click in the product page to determine the information needed by the Web users, such as authors, table of contents, and editorial reviews.
c. Use data analytics approaches, for example, decision tree, cluster algorithm, or support vector machine, to develop hypotheses about important attributes for buying a decision (Appendix 13).

5.2 Thesis Contribution

The main contribution of this thesis from a scientific perspective is providing possible adjustments for the existing data quality metrics development method. This is achieved by conducting the theory testing research strategy using a case study with new data/process domain. The possible adjustments consist of the process model configuration and the objects within the metamodel, including the process metrics, business problems and data defects matrix, data quality attributes, data quality requirements, and data quality metrics.

Another contribution of this thesis is providing a validation for the process model to integrate the data quality metrics introduced by Wang et al. [29]. Combining the two process models, we could develop the data quality metrics for the MDM environment.

The cases that have similar features—namely, book and journal data in the e-commerce process and developing a data quality metrics for product MDM—could use an alternate process model that is developed within this thesis work. While other cases with different domains are supposed to be able to implement the process model with minimal adjustments.

5.3 Research Questions

In order to meet the main objective, several research questions should be answered as follows:

i. What is the type of methodology that should be used to develop business-oriented data quality metrics?

   The selected GPF for data quality metrics development is based on the study by Otto et al. [19] and is compared in section 3.1.7 with several other data quality assessments and improvement method studies, for example, AIMQ, TIQM, DsQ-P, and ORME-DQ. Several other studies in section 3.1.7—for example, DQ metrics requirement, DQ measurements, and DQ requirements integration—remain useful during the thesis work because it is expected that the process model and the components within the metamodel are flexible enough to adjust with the selected case study. The GPF consists of three activities: identifying the business problems and data defects, specifying the DQ requirements and DQ metrics, and verification of the result. It also provides the metamodel that specifies the components needed to be developed during the process and some tangible examples of the components.

   The process model by Wang et al. [29] and Batini et al. [4] is used as the GPF for integrating data quality in the MDM environment. The main activities of the process are pre integration, comparison of the schema, conforming and merging the schemas. The study used several qualitative criteria to conform and merge the schemas, namely completeness and correctness, minimality, and understandability.

ii. How appropriate is the methodology for practical case? What processes or components should be altered?

   The feasibility of the methodology is assessed by conducting the GPF in a case study (section 4). The process models by Otto et al. [19] and Wang et al. [29] could be used in a practical case, but it should be altered to adjust the case study's environment. The alterations for the GPF cover the process model and the components in the metamodel. The alternate configuration for the process model includes the introduction of new activities and different process flows to add effectiveness to the process.

   The components in the metamodel of the general process work are sufficient for the process in Elsevier, and this thesis produces alternative objects for those components, for example, the business problems and data defects, process metrics, data quality requirements, and data quality metrics.

iii. What are the data quality metrics specifications for a study case in Elsevier?

   The data quality metrics are developed for a certain case in Elsevier, and this thesis work also evaluates the data quality metrics against the requirements and the product data. The result is a list of data quality metric specifications in Table 34 and Table 35.

The DQ metrics developed for Elsevier's case include the measurements for completeness, syntactical correctness, absence of contradiction, absence of repetition, and accuracy. Those DQ metrics are filtered using the data quality metrics requirements, database assessment, and a statistical function, which is mean and standard deviation. Because the target system is an MDM, a further activity to integrate the data quality metrics is conducted. The result is the list of integrated data quality metrics for a product MDM system. Some of the metrics developed for the study case could be used by other companies in the same industry by having a modification to adjust with their internal database structure.

5.4 Main Research Goal

In order to identify, collect, analyze, and evaluate the quality metrics for product master data, both theoretical and practical assessments should be conducted. The theoretical assessment, which aim is to select the appropriate process models, is required to provide a profound foundation for the thesis work. In the other hand, the practical assessment is required to validate the process model and to provide several components, namely the list of acceptable DQ metrics and possible adjustments for the process models in the literature studies. The adjusted GPF is feasible for the case study in Elsevier, and the developed DQ metrics meets the requirements, namely it has business impacts, it is feasible and reproducible, and it is acceptable.

The activities and components in the developed process model are complete and valid for two reasons: they are developed using studies that are validated with case studies, and they are used within the case study in Elsevier. The process model is also practical because it is feasible, reproducible, and the resulted data quality metrics meet the requirements by the users. The process model and metamodel are expected to be resilient, while the tangible objects like data quality metrics requirements and specification could be updated using new studies.

5.5 Reflection and Future Works

This thesis work answers the question on how to identify, collect, analyze, and evaluate the quality metrics for a product master data; to allow quantifying and improve their value. The answers developed in this thesis provide several results, namely a feasibility assessment and possible adjustment for a method, a list of tangible objects that are applicable to a more general case like questionnaires, data quality attributes and requirements, data quality metrics, and a list of tangible objects that are applicable for the study case like the threshold values and data structure details.

In this thesis work, we focus on the model that is developed on the basis of one research because the study of master data management is limited. Based on our assessment in section 3, there are also limited studies on data quality that focus on metrics development and provide explicit links between the business requirements and the data quality metrics. One of the reasons for using the method by Otto [19] is that it was developed using two research strategies (Verschuren et al. [27]), namely the case study strategy that uses three different cases and the grounded theory strategy that compares several methods. The three cases are customer data used in the customer service process, design data used in the manufacturing process, and material data used in the maintenance process. Basically, this thesis work is also using a combination of the two strategies to develop the altered method for developing the business-oriented data quality metrics.

Having only one case study might raise a question about its credibility to contribute for the scientific community. It is because we would assume that one cannot generalize on the basis of an individual case. Flyvbjerg [9] put this as one of the misunderstandings in a case study research and provided this statement: *"One can often generalize on the basis of a single case, and the case study may be central to scientific development via generalization as supplement or alternative to other methods. But formal generalization is overvalued as a source of scientific development, whereas 'the force of example' is underestimated."* He used several examples of falsification to point out this idea. Here, we could also use the case in Elsevier as a source of conditions for falsification to provide a new generalization; that is, the GPF might only work for certain domains.

The condition in the selected case might provide a non-ideal environment for the method, and we tried to develop the missing required components, for example, the list of KPI for e-commerce, the data quality requirements, and the data quality – performance (sales, visit) reports. Those components are considered important because the method itself needs to make a correlation between the data quality and the business performance. These conditions provide a variation in the selected case compared with the ones used to develop the GPF. Thus, we introduce some new activities or routes of action in the process model that works for product MDM in Elsevier, especially for book and journal data in the e-commerce process.

The altered process developed in this thesis is not revolutionary; that is, it still has a similar structure with the GPF. Woodal et al. [30] also got the same result when comparing several data quality methodologies to develop the hybrid model. They developed the basic process model with the ground theory research strategy using several studies. Using the case study strategy with one case, they developed a process model from the basic/generic model that fits the requirements.

On the basis of the above information, further studies are needed in order to complement this thesis work with the following objectives or features:

1. Research with a case study strategy using new domains. The goal of the research is to get a condition that could falsify the existing condition and develop a new version of the process model. The goal could also add new domains that could be satisfied by existing process models.
2. Research that studies which parts of the business process contribute the most to the process model alteration. Otto et al. [19] did not provide the characteristics of the three case studies used to develop the process model except for their domain. A study to make a correlation between the case study's features is required to provide a generic configuration between the business process type and the proper process model configuration.
3. Research that incorporates the data quality cost. This component is important when developing the KPI and the business problems because it is expected that the company does not have the financial report that can be linked directly with the data quality defect. This thesis work develops the business problems using the expert's experience and tries to make a correlation with the performance by analyzing the performance data. Providing the potential costs for a possible data defect might enrich the business problems and provide more credibility because the costs are developed using other business cases.

Appendix 1 Data Quality Process

Table 14 Process - Information Matrix, Otto et al. [19]

Information	Corporate data steward	Client	Process owner	Data users	Senior technical data steward	Technical data steward
Phase I: Identification						
I.1 Identify Business Process and Process Metrics						
Process activity. Identifier	x	x	x			
Process activity. Business Process	x	x	x			
Process activity. Accountability	x		x			
Process indicator (all)	x		x			
Business problem (all)	x		x			
Business goal (all)	x		x			
I.2 Identify Data and IT System						
Process activity. Business Application	x		x			
Data (all)	x		x			
I.3 Identify business problems and data defects						
Business goal. Impairment	x		x			
Business problems. Cause	x			x	x	x
Data defect	x			x	x	x
Preventive measure (all)	x			x	x	x
Reactive measure (all)	x			x	x	x
Phase II: Analyze and Specify						
II.1 Define and Rank Requirements for Data Quality Metrics						
Data quality indicator. Requirements	x	x			x	x
II.2 Specify Data Quality Metrics						
Data. Constraints	x				x	x
Data quality indicator. Identifier	x				x	x
Data quality indicator. Dimension	x				x	x
Data quality indicator. Measuring point	x					x
Data quality indicator. Measuring method	x					x
Data quality indicator. Scale level	x				x	x
Data quality indicator. Unit	x				x	x
Data quality indicator. Measurement frequency	x				x	x
Phase III: Verify and Document						
III.1 Verify requirement fulfillment	x				x	x
III.2 Document data quality metrics specification	x	x				

Appendix 2 Data Quality Dimensions

Table 15 Data Quality Dimension, Sebastian-Coleman [23]

No	Dimensions	Code	Definition
1	Completeness	SC-01	completeness implies having all the necessary or appropriate parts; being entire, finished, total. The first condition of completeness is existence
2	Timeliness	SC-02	▪ the degree to which data represent reality from the required point in time (English, 1999) ▪ the degree to which customers have the data they need at the right time
3	Validity	SC -03	the degree to which data conform to a set of business rules
4	Consistency	SC -04	the absence of variety or change
5	Integrity	SC -05	▪ the degree to which data conform to data relationship rules (as defined by the data model) that are intended to ensure the complete, consistent, and valid presentation of data representing the same con-cepts ▪ it is reserved for cross-table relationships
6	Accuracy	SC -06	▪ Accurate data is true and correct. ▪ approach the accuracy from validity dimension. While validity is not accuracy (valid values can be incorrect), there is still knowledge to be gained by measuring validity (invalid values cannot be correct).

Table 16 Data Quality Dimension, Morbey [17]

No	Dimensions	Code	Definition
1	Completeness per row (horizontal completeness)	G-01	Is there any missing or defective data in a record? All data is entered according to business needs.
2	Syntactical correctness (conformity)	G-02	Is there data in a non-standardized format? The data fits into the specific format
3	Absence of contradictions (consistency)	G-03	Which data values are contradictory? The data do not contradict integrity specifications (business rules, empirical values) or defined ranges of values (within the data pool, in comparison with other data pools, in time elapsed).
4	Accuracy incl. currency	G-04	Which data is wrong or expired? Correct and up to date (timeliness) notation of existing names, addresses, products etc.
5	Absence of repetitions (free of duplicates)	G-05	Which data records or contents of columns are being repeated? No duplicates (search for synonyms and similarities), no homonyms, no overlapping (continuity), everything is precisely identifiable (uniqueness).
6	Business referential integrity (integrity)	G-06	Which reference data or relations are missing? There will not be any clients without a contract, products will be listed.
7	Completeness (Cross check sums, vertical completeness)	G-07	Is there data consistency over all systems?
8	Availability of documentation (findability)	G-08	Can the data be found easily and quickly (e.g. using common “search”-functions)
9	Normative consistency	G-09	It has to be assured that the naming and meaning of certain data is the same over all systems, processes and departments of the organization.

Table 17 Data Quality Dimensions by Batini et al. [2]

No	Dimensions	Code	Definition
1	Accuracy	B-01	closeness between a value of v and a value of v', considered as the correct representation of the real-life phenomenon that v aims to represent
a	Syntactic accuracy	B-01a	the closeness of a value v to the elements of the corresponding definition domain D
b	Semantic accuracy	B-01b	the closeness of the value v to the true value v'
2	Completeness	B-02	the extent to which data are of sufficient breadth, depth, and scope for the task at hand
a	Schema Completeness	B-02a	the degree to which concepts and their properties are not missing from the schema
b	Column Completeness	B-02b	a measure of the missing values for a specific property or column in a table
c	Population Completeness	B-02c	evaluates missing values with respect to a reference population
3	Time related dimensions		
a	Currency	B-03a	concerns how promptly data are updated
b	Volatility	B-03b	characterizes the frequency with which data vary in time
c	Timeliness	B-03c	expresses how current data are for the task at hand
4	Consistency	B-04	capture the violation of semantic rules defined over (a set of) data items, where items can be tuples of relational tables or records in a file.
5	Others		
a	Accessibility	B-05a	the ability of the user to access the data from his or her own culture, physical status/functions, and technologies available.
b	Quality of Information sources	B-05b	this dimension could be a composition of believability, reputation, objectivity, and reliability (credibility) dimensions of Wang's

Table 18 Data Quality Dimensions by Zang

Dimension	Description
Completeness	Are all necessary data present or missing?
Validity	Are all data values within the valid domains specified by the business?
Integrity	Are the relations between entities and attributes consistent?
Duplication	Are there multiple, unnecessary representations of the same data objects?
Consistency	Is data consistent between systems?
Timeliness	Is data available at the time needed?
Accuracy	Does data reflect the real world objects or a verifiable source?

Appendix 3 Data Quality Measurements

Table 19 Data Quality Measurements by Batini et al. [3]

Dimensions	Name	Metrics Definition
Accuracy	Acc1	Syntactic accuracy: it is measured as the distance between the value stored in the database and the correct one. Syntactic Accuracy=Number of correct values/number of total values
	Acc2	Number of delivered accurate tuples
	Acc3	User Survey - Questionnaire
Completeness	Compl1	Completeness = Number of not null values/total number of values
	Compl2	Completeness = Number of tuples delivered/Expected number
	Compl3	Completeness of Web data = $(T_{max}- T_{current})^{*}$ $(Completeness_{Max}- Completeness_{Current})/2$
Consistency	Cons1	Consistency = Number of consistent values/number of total values
	Cons2	Number of tuples violating constraints, number of coding differences
Timeliness	Time1	Timeliness = $(\max (0; 1\text{-Currency/Volatility}))^{s}$
	Time2	Percentage of process executions able to be conducted within the required time frame
	Time3	User Survey - Questionnaire
Currency	Curr1	Currency = Time in which data are stored in the system - time in which data are updated in the real world
	Curr4	Currency = Age + (Delivery time- Input time)
	Curr5	User Survey - Questionnaire
Volatility	Vol1	Time length for which data remain valid
Uniqueness	Uni1	Number of duplicates
Appropriate amount of data	Appr1	Appropriate Amount of data = Min ((Number of data units provided/Number of data units needed); (Number of data units needed/Number of data units provided))
	Appr2	User Survey - Questionnaire
Accessibility	Access1	Accessibility = max (0; 1-(Delivery time - Request time)/(Deadline time - Request time))
	Access3	User Survey - Questionnaire
Credibility	Cred1	Number of tuples with default values
	Cred2	User Survey - Questionnaire
Interpretability	Inter1	Number of tuples with interpretable data, documentation for key values
	Inter2	User Survey - Questionnaire
Usability	Usa1	User Survey - Questionnaire
Conciseness	Conc1	Number of deep (highly hierarchic) pages
	Conc2	User Survey - Questionnaire
Maintainability	Main1	Number of pages with missing meta-information
Applicability	App1	Number of orphaned pages
	App2	User Survey - Questionnaire
Convenience	Conv1	Difficult navigation paths: number of lost/interrupted navigation trails
Speed	Speed1	Server and network response time
Comprehensiveness	Comp1	User Survey - Questionnaire
Clarity	Clar1	User Survey - Questionnaire
Traceability	Trac1	Number of pages without author or source
Security	Sec1	Number of weak log-ins

Dimensions	Name	Metrics Definition
	Sec2	User Survey - Questionnaire
Correctness	Corr1	User Survey - Questionnaire
Objectivity	Obj1	User Survey - Questionnaire
Relevancy	Rel1	User Survey - Questionnaire
Reputation	Rep1	User Survey - Questionnaire
Ease of operation	Ease1	User Survey - Questionnaire

Table 20 Examples of Data Quality Measurements, Sebastian-Coleman [23]

Dimension of Quality	Measurement Type	Measurement Type Description	Assessment Category
Timeliness	Timely delivery of data for processing	Compare actual time of data delivery to scheduled data delivery	In-line measurement
Completeness	Field completeness - non-nullable fields	Ensure all non-nullable fields are populated	Process control
Integrity/ Completeness	Dataset integrity - duplicate record reasonability check	Reasonability check, compare ratio of duplicate records to total records in a dataset to the ratio in previous instances of dataset	In-line measurement
Timeliness	Timely availability of data for access	Compare actual time data is available for data consumers access to scheduled time of data availability	In-line measurement
Validity	Validity check, single field, detailed results	Compare values on incoming data to valid values in a defined domain (reference table, range, or mathematical rule)	In-line measurement
Validity	Validity check, roll-up	Summarize results of detailed validity check; compare roll-up counts and percentage of valid/invalid values to historical levels	In-line measurement
Integrity/ Validity	Validity check, multiple columns within a table, detailed results	Compare values in related columns on the same table to values in a mapped relationship or business rule	In-line measurement
Consistency	Consistent column profile	Reasonability check, compare record count distribution of values (column profile) to past instances of data populating the same field.	In-line measurement
Consistency	Consistent dataset content, distinct count of represented entity, with ratios to record counts	Reasonability check, compare distinct counts of entities represented within a dataset (e.g., the distinct number of customers represented in sales data) to threshold, historical counts, or total records	In-line measurement

Dimension of Quality	Measurement Type	Measurement Type Description	Assessment Category
Consistency	Consistent dataset content, ratio of distinct counts of two represented entities	Reasonability check, compare ratio between distinct counts of important fields/entities (e.g., customers/sales office, claims/insured person) to threshold or historical ratio	In-line measurement
Consistency	Consistent multicolumn profile	Reasonability check, compare record count distribution of values across multiple fields to historical percentages, in order to test business rules (multicolumn profile with qualifiers)	In-line measurement
Consistency	Consistent record counts by aggregated date	Reasonability check, compare record counts and percentage of record counts associated an aggregated date, such as a month, quarter, or year, to historical counts and percentages	Periodic measurement
Integrity/ Completeness	Parent/child referential integrity	Confirm referential integrity between parent/child tables to identify parentless child (i.e., orphan) records and values	Periodic measurement
Integrity/ Completeness	Child/parent referential integrity	Confirm referential integrity between child/parent tables to identify childless parent records and values	Periodic measurement
Integrity/ Validity	Validity check, cross table, detailed results	Compare values in a mapped or business rule relationship across tables to ensure data is associated consistently	Periodic measurement
Integrity/ Consistency	Consistent cross-table multicolumn profile	Cross-table reasonability check, compare record count distribution of values across fields on related tables to historical percentages, in order to test adherence to business rules (multicolumn profile with qualifiers)	Periodic measurement
Consistency	Consistency compared to external benchmarks	Compare data quality measurement results to a set of benchmarks, such external industry or nationally established measurements for similar data	Periodic measurement

Appendix 4 Business Problems and Data Defects in E-commerce

a. Business Problem

The company sells its products through several Web-based channels. The e-commerce site within this context is an e-store (store.elsevier.com) where the product data is managed in a single repository. The e-commerce website sells books and journals in print and electronic formats. E-commerce offers low barriers for potential customers to access; thus it is expected to increase the sales.

The common performance elevation expected through e-commerce are lower time to market, increase in sales, and lowering in cost; better customer satisfaction through better accessibility, speed, and higher visibility; and agility in business to adjust to customers' needs. Those are also the challenges in e-commerce. The poor quality of product information could inhibit the performance expectation. The business problems that are caused by poor data in e-commerce are described in Table 21.

Table 21 Causal Relation: Business Problem and Data Defect

No	Business Problem	Business Impact	Data Defect	DQ Dimensions	Attribute
i	Customer does not buy a product	Potential revenue loss	Incomplete information in the e-commerce system (book)	Completeness per row	All in websites data model
ii	Customer could not browse the site conveniently	Customer dissatisfaction	Taxonomy mapping problem	Absence of contradiction	Subject/ Category
iii	Unable to run a marketing campaign using AdWords and other e-mail channels	Potential revenue loss	Incomplete information in the e-commerce system	Completeness per row	All in marketing data model
iv	Internet user could not find the data in the top result using a search engine	Potential revenue loss	Incomplete information in the e-commerce system	Completeness per row	All in websites data model
v	Offering an unavailable product	Customer dissatisfaction, unrecognized revenue, ineffective marketing, and potential revenue loss	Inaccurate data in the e-commerce system (journal)	Accuracy including currency	Saleable/ Availability in a region
			Incomplete data in the e-commerce system (journal)	Completeness, business referential integrity	Fulfillment system
			Inconsistent data between the journal database and the e-commerce system	Absence of contradiction, accuracy including	Product data

No	Business Problem	Business Impact	Data Defect	DQ Dimensions	Attribute
			Inaccurate data in the e-commerce system	currency	Product data
vi	Products are not included in the marketing campaign	Potential revenue loss	Taxonomy mapping problem	Absence of contradiction	Subject/ Category

b. Reactive and Preventive Measures

The need for high-quality product data is considered important in e-commerce. It is needed to provide sufficient and correct information about a book or a journal to the potential buyer, to ensure that they will buy the right product or the product has the content they need. There are several activities to maintain product data quality within the e-commerce system in Elsevier, which are described in Table 22.

Table 22 Preventive and Reactive Measures

No	Type	Type	DQ Dimensions	Attribute	Actor
i	DQ check at data import	Preventive	Currency, business referential integrity, absence of repetition	All in repository data model	ETL tool
ii	Manual update using the e-commerce system	Reactive	Completeness per row, business referential integrity, accuracy incl. currency, absence of contradictions, absence of repetitions	All in website data model	Marketing/ Sales staff

Appendix 5 Requirement for Data Quality Metrics

Table 23 DQ Metrics Requirements, DAMA

No	Requirement	Description
1	Measurability	A data quality metric must be measurable, and should be quantifiable within a discrete range
2	Business Relevance	The value of the metric is limited if it cannot be related to some aspect of business operations or performance. Therefore, every data quality metric should demonstrate how meeting its acceptability threshold correlates with business expectations
3	Acceptability	The data quality dimensions frame the business requirements for data quality, and quantifying quality measurements along the identified dimension provides hard evidence of data quality levels. Base the determination of whether the quality of data meets business expectations on specified acceptability thresholds
4	Accountability / Stewardship	Associated with defined roles indicating notification of the appropriate individuals when the measurement for the metric indicates that the quality does not meet expectations
5	Controllability	Any measurable characteristic of information that is suitable as a metric should reflect some controllable aspect of the business. In other words, the assessment of the data quality metric's value within an undesirable range should trigger some action to improve the data being measured.
6	Trackability	Quantifiable metrics enable an organization to measure data quality improvement over time. Tracking helps data stewards monitor activities within the scope of data quality SLAs, and demonstrates the effectiveness of improvement activities

Table 24 DQ Metric Requirements, Heinrich [11]

No	Requirement	Description
	representation consistency	
1	Normalization	Assure that the values of the metrics are comparable.
2	Interval Scale	The metrics are in interval scale.
3	Interpretability	The DQ metrics have to be comprehensible. E.g., considering a metric for timeliness, it could be interpretable as the probability that a given attribute value within the database is still up-to-date
	interpretation consistency and *aggregation consistency*	
4	Aggregation	the metrics must allow aggregation of values on a given level to the next higher level
	impartial-contextual consistency	
5	Adaptivity	The metrics can be adapted to the context of a particular application
	additional	
6	Feasibility	When defining metrics, measurement methods should be defined and in cases when exact measurement is not possible or cost-intensive, alternative (rigorous) methods (e.g. statistical) shall be proposed.

Table 25 Characteristics of Effective Measurement, Sebastian-Coleman [23]

No	Requirements	Description
1	Measurements must be Comprehensible and Interpretable	To be effective, measurements themselves must be comprehensible. If people cannot understand what characteristic is being measured, the measurement will not help reduce uncertainty or be useful, even if the object being measured is very

No	Requirements	Description
		important. An example is in using a thermometer, we understand what is to measure, understand the scale and how to read the result, and understand the threshold to make a decision.
2	Measurements must be Reproducible	The main reason for focusing on the instruments of measurement (rulers, scales, and the like) and the conditions of measurement (temperature, age, etc.) is to produce consistent measurement results and to understand any factors that might introduce variability into the measurement
3	Measurements must be Purposeful	We need to have a reason for measuring the things we measure. Businesses have developed financial and performance measurements in order to make decisions about what skills to look for in employees, where to make long term investments, and how to prepare for future opportunities.

Table 26 DQ Metrics Requirements, Huner [12]

No	Requirements	Description
1	Understandability and Complete Information	A DQ Metrics should have metadata to provide correct interpretation of their value and describe its purpose. A DQ Metrics should have this information in the metadata: measuring frequency, measuring point, measurement method, scale, threshold/ target value, escalation process, the data items, and who is accountable.
2	Relation with other components	A DQ Metrics should be assigned to one or more DQ dimensions, process metrics, strategic objective, and business problem
3	Acceptability	See DAMA in Table 1
4	Controllability	
5	Business Relevance	
6	Measurability	
7	Normalization	See Kaiser in Table 2
8	Aggregation	
9	Cost/ Benefit	The effort required for the definition and collection of the values of a DQ measure should be justified by the benefits (controlled potential for error).
10	SMART principles	A DQ metrics should meet the SMART goals (specific, measurable, attainable, relevant, and time-bound).
11	Comparability	
12	Use in SLAs	A DQ metrics should be able to be used in service level agreements.
13	Visualization	The values of a DQ metrics should be able to be visualized (e.g. time series graphs).
14	Repeatability	Values for a DQ metrics should be applicable not only once but several times.

Table 27 Data Quality Requirements, Loshin [15]

No	Requirements	Description
1	Clarity of Definition	Explains what is being measured, the key stakeholders participate in its definition and agree to the definition's final wording, advisable to provide the metric's value range, as well as a qualitative segmentation of the value range that relates the metric's score to its performance assessment.
2	Measurability	See DAMA
3	Business Relevance	See DAMA. More desirable is if that performance measurement can be directly associated with a critical business impact
4	Controllability	See DAMA
5	Representation	One should associate a visual representation that logically presents the metric's value in a concise and meaningful way.

6	Reportability	Each metric's definition should provide enough information that can be summarized as a line item in a comprehensive report. The difference between representation and reportability is that the representation will focus on the specific metric in isolation, whereas the reporting should show each metric's contribution to an aggregate assessment.
7	Trackability	See DAMA
8	Drill-Down Capability	The ability to expose the underlying data that contributed to a particular metric score

Appendix 6 eCommerce Metrics

Table 28 eCommerce KPI, Tsai et al. [26]

KPIs construct	KPIs	Mean ± S.D.	CV %	Quartile deviation	Median
Financial	Service cost	8.50 ± 0.69	8.12	0.50	9.00
	Financial earning	8.50 ± 0.69	8.12	0.50	9.00
	Appropriate budget control	7.75 ± 0.44	5.68	0.13	8.00
	Sales growth rate	7.10 ± 1.07	15.07	1.00	7.00
	Market share	7.75 ± 0.91	11.74	0.50	8.00
Customer	Willingness to purchase	8.60 ± 0.68	7.91	0.50	9.00
	Customer satisfaction	7.40 ± 0.99	13.38	0.50	7.00
	Product information	7.40 ± 0.60	8.11	0.50	7.00
	Increase in trust from customers	7.85 ± 0.59	7.52	0.13	8.00
	Search engine optimization	7.50 ± 0.89	11.87	0.50	7.00
	Convenience in product ordering	7.50 ± 0.83	11.07	0.50	7.00
	Payment function	8.50 ± 0.76	8.94	0.50	9.00
	Rapid delivery	8.55 ± 0.83	9.71	0.13	9.00
	After-sales service	7.60 ± 0.94	12.37	0.50	7.50
Internal process	Efficiency in managing orders	7.25 ± 1.07	14.76	0.50	8.00
	Function of the information system	7.35 ± 0.88	11.97	0.50	7.00
	Ability to write marketing proposals	7.55 ± 0.69	9.14	0.50	7.00
	Ability to conduct internet marketing	8.50 ± 0.76	8.94	0.50	9.00
	Selection of products for display	8.40 ± 0.75	8.93	0.50	9.00
	Customer complaint management	7.40 ± 1.35	18.24	0.50	8.00
	Transaction safety and assurance	8.40 ± 0.75	8.93	0.50	9.00
	Innovative service process	7.40 ± 0.99	13.38	0.50	7.00
Learning and growth	Employee's willingness to learn	8.50 ± 0.76	8.94	0.50	9.00
	Employee training programs	7.70 ± 0.80	10.39	0.50	8.00
	Employee's ability to conduct Internet marketing	8.55 ± 0.83	9.71	0.13	9.00
	Efficiency of teamwork	6.45 ± 1.00	15.50	0.50	7.00
	Knowledge sharing culture	7.45 ± 0.83	11.14	0.50	7.00
	Employee satisfaction	8.15 ± 0.59	7.24	0.13	8.00
	Application of market information	7.40 ± 0.88	11.89	0.50	7.00

Appendix 7 Phase I: Business Problems and Poor Data in Elsevier eCommerce

Table 29 Simple Assessment Result on Marketing Data

No	Defect	Items Affected			Dimensions
		Records	% Records	USD	
a	BLANK ISBN	-	-	0	Completeness
b	BLANK TITLE	-	-	0	Completeness
c	BLANK SUBTITLE	24,554	67.32	4,175,581.14	Completeness
d	BLANK OVERVIEW	6,464	17.72	1,036,709.75	Completeness
e	UNKNOWN AUTHOR	2,550	6.99	500,464.12	Syntactical Correctness
f	ERROR IMAGE	6,105	16.74	848,764.51	Accuracy
g	Available in EU but error*	294	0.81	37,375.34	Accuracy
h	Not Available in EU but success	451	1.24	80,311.34	Accuracy

*54 products do not have price in EUR

Appendix 8 Data Quality Metrics Attributes

Table 30 Data Quality Metrics Attributes, Huner [12]

No	Attributes	Description
1	Identifiers	Identifiers for a DQ Metrics
2	Dimension	Related dimension e.g. accuracy, completeness
3	Measuring point	Where the measurement activity takes place e.g. product database
4	Measurement method	Methods to use for measurement. An example is to count number of complete rows divided to all rows to compute completeness
5	Scale level	The scale that is used e.g. interval, ratio
6	Unit	The unit for the value e.g. percentage, rows
7	Measurement frequency	Frequency of measurement activity e.g. daily, weekly
8	Requirements	Which DQ requirements are met
9	Data	Which data (part of data) is relevant? This attribute also gives information about related business process
10	Data Defect	Which data defect is this metrics for? This attribute also gives information about related business problem

Table 31 Data Quality Metrics Attributes, Sebastian-Coleman [23]

No	Attribute Name	Attribute Definition
1	Measurement Type Number	This field identifies the DQAF measurement type.
2	Specific Metric Number	This field contains a unique number that serves as a key
3	Dataset Name	This field contains the name of the dataset being measured.
4	Dataset Source	This field contains the name of the source system
5	Dataset Type	This field refers to the form the dataset takes. For example, a dataset can be a file, a set of messages, or a table.
6	Range Minimum	For any measurement on the basis of a range of values, this field represents the lowest value in that range.
7	Range Maximum	For any measurement on the basis of a range of values, this field represents the highest value in that range.
8	Data Quality Threshold Type	Valid values for Data Quality Threshold Type are: manual, automated on the basis of mean, automated on the basis of median, automated on the basis of average.
9	Data Quality Threshold (if threshold is set manually)	This field contains the data quality threshold number for thresholds that are set manually. This field is populated only if the type is manual.
	Optional	
10	Metric Name	Provides a name that enables people to understand what data is being measured.
11	Metric Description	Provides additional information needed to understand what data is being measured.
12	Metric Criticality	Records a level of criticality for the metric; for example, high, medium, low.
13	Business Contact/Steward	Provides the name of a businessperson who needs to be informed if the metric generates an unusual result.
14	Date the measurement was established	Records the Effective Date or Start Date for the metric

No	Attribute Name	Attribute Definition
15	Date the measurement was made inactive	Records the End Date or Expiration Date for the metric
16	Active Indicator	Shows whether the metric is active; prevents the collection of additional results if there is a need to turn the metric off.
17	Frequency at which measurement should be executed	Describes how often the metric should be run.
18	Notification indicator	Records whether a notification should be sent if a metric produces an unusual result
19	Notification contact person	Records the name of the person who should be contacted if the measurement produces an unusual result.
20	Notification contact information	For automated notifications, the contact information is likely to be an e-mail address.
21	Denominator Field	Field that has complex type like a table. This could be useful for metrics that uses historical value distribution

Table 32 Developed DQ Metrics Attributes, Hüner et al. [12] and Sebastian-Coleman [23]

No	Attributes	Description
1	Identifiers	Identifiers for a DQ Metrics
2	Dimension	Related dimension e.g. accuracy, completeness
3	Measuring point	Where the measurement activity takes place e.g. product database
4	Measurement method	Methods to use for measurement. An example is to count number of complete rows divided to all rows to compute completeness
5	Scale level	The scale that is used e.g. interval, ratio
6	Unit	The unit for the value e.g. percentage, rows
7	Measurement frequency	Frequency of measurement activity e.g. daily, weekly
8	Requirements	Which DQ requirements are met
9	Data	Which data (part of data) is relevant? This attribute also gives information about related business process
10	Data Defect	Which data defect is this metrics for? This attribute also gives information about related business problem
11	Threshold	The data quality threshold number

Appendix 9 Measurement Methods for eCommerce

Table 33 Data Quality Measurement Definition

Dimension	Study	Definition of Measurement
Completeness per Row	DAMA	Assign completeness rules to a data set in varying levels of constraint–mandatory attributes that require a value, data elements with conditionally optional values, and inapplicable attribute values
	Coleman	**Field completeness - non-nullable fields**: Ensure all non-nullable fields are populated
	Coleman	**Parent/child referential integrity**: Confirm referential integrity between parent/child tables to identify parentless child (i.e., orphan) records and values (Integrity/Completeness)
	Coleman	**Child/parent referential integrity**: Confirm referential integrity between child/parent tables to identify childless parent records and values (Integrity/Completeness)
Syntactical Correctness (Conformity)	Peralta	**Syntactic Correctness Ratio Metric:** The most typical syntactical rules checks for illegal values (e.g. out-of-range), non-standard format or embedded values (e.g. “Paris, France” in a city attribute). Note that when evaluating semantic correctness metrics, when the tuple is a mismember, all the attribute values are inaccurate and when the key is not accurate most of the attribute values are inaccurate (except for hazard coincidences).
	Peralta	**Syntactic Correctness Deviation Metric:** It measures the syntactic distance between a system datum and some neighbor data that is syntactically correct. As an example, consider the Name attribute of Table 7 as a reference catalog for students’ names. The nearest element for “A. Benedetti” is “Ana Benedetti” and the value-deviation metric can be calculated using some edit distance function
Absence of Contradiction (Consistency) and Normative Consistency	DAMA	A set of rules that specify consistency relationships between values of attributes, either across a record or message, or along all values of a single attribute
	Coleman	**Consistent column profile**: Reasonability check, compare record count distribution of values (column profile) to past instances of data populating the same field.
	Coleman	**Consistent dataset content, distinct count of represented entity, with ratios to record counts**: Reasonability check, compare distinct counts of entities represented within a dataset (e.g., the distinct number of customers represented in sales data) to threshold, historical counts, or total records
	Coleman	**Consistent dataset content, ratio of distinct counts of two represented entities**: Reasonability check, compare ratio between distinct counts of important fields/entities (e.g., customers/sales office, claims/insured person) to threshold or historical ratio
	Coleman	**Consistent multi columns profile**: Reasonability check, compare record count distribution of values across multiple fields to historical percentages, in order to test business rules
	Coleman	**Consistent amount field calculations across secondary fields:** Reasonability check, compare amount column calculations, sum (total) amount, percentage of total amount, and average amount across a secondary field or fields to historical counts and percentages, with qualifiers to narrow results.
	Coleman	**Consistent record counts by aggregated date**: Reasonability check, compare record counts and percentage of record counts associated an aggregated date, such as a month, quarter, or year, to historical counts and percentages

Dimension	Study	Definition of Measurement
	Coleman	**Consistent amount field data by aggregated date**: Reasonability check, compare amount field data (total amount, percentage of total amount) aggregated by date (month, quarter, or year) to historical total and percentage
	Coleman	**Consistent cross-table multicolumn profile**: Cross-table reasonability check, compare record count distribution of values across fields on related tables to historical percentages, in order to test adherence to business rules (Integrity/Consistency)
Absence of Repetitions (Free of Duplicates)	DAMA	no entity exists more than once within the data set and that a key value relates to each unique entity, and only that specific entity, within the data set
Business referential integrity (Integrity)	DAMA	specifying that when a unique identifier appears as a foreign key, the record to which that key refers actually exists
	Coleman	**Parent/child referential integrity**: Confirm referential integrity between parent/child tables to identify parentless child (i.e., orphan) records and values (Integrity/Completeness)
	Coleman	**Child/parent referential integrity**: Confirm referential integrity between child/parent tables to identify childless parent records and values (Integrity/Completeness)
	Coleman	**Consistent cross-table multicolumn profile**: Cross-table reasonability check, compare record count distribution of values across fields on related tables to historical percentages, in order to test adherence to business rules (Integrity/Consistency)
	Coleman	**Validity check, multiple columns within a table, detailed results**: Compare values in related columns on the same table to values in a mapped relationship or business rule (Integrity/Validity)
	Coleman	**Validity check, cross table, detailed results**: Compare values in a mapped or business rule relationship across tables to ensure data is associated consistently (Integrity/Validity)
Vertical completeness	Coleman	**Consistent cross-table multicolumn profile**: Cross-table reasonability check, compare record count distribution of values across fields on related tables to historical percentages, in order to test adherence to business rules (Integrity/Consistency)
Accuracy Accuracy	DAMA	Measure accuracy by how the values agree with an identified reference source of correct information, such as comparing values against a database of record or a similar corroborative set of data values from another table, checking against dynamically computed values, or perhaps applying a manual process to check value accuracy
	Coleman	**Validity check, single field, detailed results**: Compare values on incoming data to valid values in a defined domain (reference table, range, or mathematical rule)
	Coleman	**Validity check, multiple columns within a table, detailed results**: Compare values in related columns on the same table to values in a mapped relationship or business rule (Integrity/Validity)
	Coleman	**Validity check, cross table, detailed results**: Compare values in a mapped or business rule relationship across tables to ensure data is associated consistently (Integrity/Validity)
	Coleman	**Validity check, single field, detailed results**: Compare values on incoming data to valid values in a defined domain (reference table, range, or mathematical rule)

Dimension	Study	Definition of Measurement
	Coleman	**Validity check, roll-up**: Summarize results of detailed validity check; compare roll-up counts and percentage of valid/invalid values to historical levels
	Peralta	**Semantic Correctness Ratio Metric:** In practice, comparing all data against real world may be not viable, so correctness-ratio is commonly estimated via sampling. Additional metrics have been defined to measure special cases of inaccuracies: ▪ Mismembership ratio metric: It measures the percentage of mismembers, i.e., the percentage of system data without correspondent in real-world. ▪ Value inaccuracy ratio metric: It measures the percentage of system data containing errors in some attributes values or containing null values**. [Note: see Sebastian-Coleman definition with validity]**
	Peralta	**Semantic Correctness Degree Metric:** In order to compute the correctness degree of a set of elements, weighted average is typically used as an aggregation function, assigning different weights to the attributes according to their relative importance [Laboisse 2005].
	Peralta	**Semantic Correctness Deviation Metric:** In practice, if a comparison with real-world data is not possible, the comparison is done against a reference value which can be obtained from other data source or synthesized from several source values, e.g. using statistics of appearance frequency or taking an average value. **[Note: see Sebastian-Coleman definition with validity]**
	Peralta	**Syntactic Correctness: See Correctness (conformity)**
Timeliness	DAMA	measure one aspect of timeliness as the time between when information is expected and when it is readily available for use
	Coleman	**Timely delivery of data for processing**: Compare actual time of data delivery to scheduled data delivery
	Coleman	**Timely availability of data for access**: Compare actual time data is available for data consumers access to scheduled time of data availability

Appendix 10 Initial Metrics Specification for each DQ Dimension for Elsevier

1. Completeness per row (horizontal completeness)

a. Metric 1

No	Attributes	Description
1	Identifier	CPR-01
2	Dimension	Completeness per Row, (in)Accuracy
3	Measurement method	**result** = 1 - (Number of row with empty non-null able field divided with number of all row)
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Sebastian-Coleman (Field completeness - non-null able fields), DAMA, Peralta (Semantic Correctness Ratio Metric**)**
	Attribute type	String and Numeric

b. Metric 2

No	Attributes	Description
1	Identifier	CPR-02
2	Dimension	Completeness per Row, Business Referential Integrity (Integrity)
3	Measurement method	**result** = 1 - (Number of unreferenced row (parentless row) divided with number of all rows)
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Sebastian-Coleman (Parent/child referential integrity)
	Attribute Type	String and Numeric

c. Metric 3

No	Attributes	Description
1	Identifier	CPR-03
2	Dimension	Completeness per Row, Business Referential Integrity (Integrity)
3	Measurement method	**result** = 1 - (Number of row with empty non-null able reference field and non-exist reference field value (childless row) divided with number of all rows)
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
8	Definition	Sebastian-Coleman (Child/parent referential integrity)
9	Attribute Type	String and Numeric

d. Metric Average

No	Attributes	Description
1	Identifier	CPR-04

2	Dimension	Syntactical correctness (conformity)
3	Measurement method	**result =** Average (CPR-01, CPR-02, CPR-03)
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
8	Definition	Coleman, DAMA
9	Attribute Type	String and Numeric

2. Syntactical correctness (conformity)

a. Metric 1

No	Attributes	Description
1	Identifier	SC-01
2	Dimension	Syntactical correctness (conformity), Accuracy
3	Measurement method	**result =** 1 - (Number of row with non-standard value or format divided with number of all rows) ▪ Standard Format: Top-3 string pattern on the basis of distribution OR defined business rule (postcode is 4 char, dash, 2 numeric: ZZZZ-99) ▪ Standard Value: Top-3 value on the basis of distribution OR between min-max value of previous data OR defined business rule (price is >=0)
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Sebastian-Coleman (Syntactic Correctness Ratio Metric; Validity check, single field, detailed results)
	Attribute Type	▪ Numeric: between min-max value ▪ String and Numeric: business rule, string patter, top-3 value

b. Metric 2

No	Attributes	Description
1	Identifier	SC-02
2	Dimension	Syntactical correctness (conformity), Normative Consistency, Accuracy
3	Measurement method	**result =** 1 - (Number of row with deviated value divided with number of all rows) ▪ Non-deviated value: there is a similar value at reference table with similarity>=0.8 for example (Levenshtein distance/length of longer string) <=0.2 OR Jaro-Winkler distance>=0.8. ▪ Similarity=1 for numeric type field
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Sebastian-Coleman (Syntactic Correctness Deviation Metric; Validity check, single field, detailed results:)
	Attribute Type	String and Numeric (deviation=0)

c. Metric Average

No	Attributes	Description
1	Identifier	SC-03
2	Dimension	Syntactical correctness (conformity)
3	Measurement method	**result =** Average (SC-01, SC-02) if the reference table for SC-02 is not available then SC-03 = SC-01
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Coleman
	Attribute Type	String and Numeric

3. Absence of contradictions (consistency) and normative consistency [v]

a. Metric 1

No	Attributes	Description
1	Identifier	AOC-01
2	Dimension	Absence of contradictions (consistency) and normative consistency
3	Measurement method	**result =** 1 - (number of non-reasonable fields divided with number of all fields) ▪ Reasonable field: field that has the same top-5 values on the basis of its distribution compared with previous data
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Sebastian-Coleman (Consistent column profile)
	Attribute Type	String and Numeric

b. Metric 2

No	Attributes	Description
1	Identifier	AOC-02
2	Dimension	Absence of contradictions (consistency) and normative consistency
3	Measurement method	▪ An example for availability - price relation ▪ [Availabilities i - PRICE i']: [EU, UK - EUR]; [US, EMEA, ASIA - USD]; [AU - AUD]; [JPY - YEN] ▪ Ratio per avail (i) = num row Avail(i) and Price (i') / num row Avail(i) ▪ **result =** average (all ratio per avail)
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Monthly, Quarterly
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Sebastian-Coleman (Consistent dataset content, distinct count of represented entity, with ratios to record counts; Consistent cross table multi columns profile:)
	Attribute Type	String and Numeric

c. Metric 3

No	Attributes	Description
1	Identifier	AOC-03
2	Dimension	Absence of contradictions (consistency) and normative consistency
3	Measurement method	▪ val1: (M-1 rows/M-2 rows); val2: (last year M-1 rows/ M-2 rows) ▪ val3 = val1/ val2 ▪ minVal = min(val1,val2,val3); maxVal=max(val1,val2,val3) ▪ rawVal = not (minVal or maxVal) ▪ **result =** (rawVal-minVal) / (maxVal-minVal) ▪ Quarterly: change M with Q
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Monthly, Quarterly
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Sebastian-Coleman (Consistent record counts by aggregated date)
	Attribute Type	Row

d. Metric Average

No	Attributes	Description
1	Identifier	AOC-03
2	Dimension	Absence of contradictions (consistency) and normative consistency
3	Measurement method	**result =** Average (SC-02, AOC-01, AOC-02, AOC-3)
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	DAMA, Coleman
	Attribute Type	String and Numeric

4. Absence of repetitions (free of duplicates)

a. Metric 1

No	Attributes	Description
1	Identifier	AOR-01
2	Dimension	Absence of repetitions (free of duplicates)
3	Measurement method	**result =** 1 - (Number of duplicate row divided with number of all unique rows) ▪ Unique = unique ISBN for book or unique ISSN for journal
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	DAMA
	Attribute Type	String and Numeric

5. Vertical Completeness

See ACR-01.

6. Business referential integrity (integrity)

a. Metric Average

No	Attributes	Description
1	Identifier	BRI-01
2	Dimension	Business Referential Integrity
3	Measurement method	**result =** Average (CPR-02, CPR-03, AOC-02)
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	DAMA, Coleman
	Attribute Type	String and Numeric

7. Accuracy incl. currency

a. Metric 1

No	Attributes	Description
1	Identifier	ACR-01
2	Dimension	Accuracy
3	Measurement method	**Result** = average(number of unique book ISBN in Book database + number of unique book ISBN in French site XML/ number of unique book ISBN in eCommerce system, number of unique ISSN in Journal database/ number of unique ISSN in eCommerce system) ▪ minVale/ maxValue
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Peralta (Semantic Correctness Ratio Metric), DAMA. Book database, Journal database, and Frenc Site XML is considered as the "Real World"
	Attribute Type	String and Numeric

b. Metric 2

No	Attributes	Description
1	Identifier	ACR-02
2	Dimension	Timeliness
3	Measurement method	▪ Ratio 1:1- (min difference of time data in Journal database/Book database/French XML with time data in eCommerce system)/ 720 ▪ Ratio 2:1- (min difference of time data in eCommerce system and time data in Web)/ 720 ▪ **result** = average(Ratio 1, Ratio 2) ▪ 720 minutes = 12 hours, if difference>720 then Ratio (i) = 0
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0

6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	DAMA, Sebastian-Coleman (Timely delivery of data for processing, Timely availability of data for access)
	Attribute Type	Numeric

c. Metric Average

No	Attributes	Description
1	Identifier	ACR-03
2	Dimension	Accuracy
3	Measurement method	**Result** = number of row with exact same values of attributes from source system divided with number of all rows
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Peralta, Coleman, DAMA Book database, Journal database, and Frenc Site XML is considered as the "Real World"
	Attribute Type	String and Numeric

d. Metric Average

No	Attributes	Description
1	Identifier	ACR-04
2	Dimension	Accuracy
3	Measurement method	**Result** = average(CPR-01, SC-01, SC-02, ACR-01, ACR-02)
4	Scale level	Simple Ratio
5	Unit	Normalized, Decimal number between 0 and 1. Best=1.0
6	Measurement frequency	Daily
7	Requirements	DQ-R-01, DQ-R-05, DQ-R-06, DQ-R-07, DQ-R-08
	Definition	Peralta, Coleman
	Attribute Type	String and Numeric

Appendix 11 Data Quality Metrics Specification for eCommerce

Measuring Point : eCommerce Database [Attribute 3]

Table 34 Metrics Specification for Business Problems

No	Business Problem	Business Impact	Data Defect	DQ Dimensions	Attribute
i	Customer does not buy a product	Potential revenue loss	Incomplete information in eCommerce system (Book database) [Attribute 10]	Completeness per row [Attribute 2]	All in websites data model [Attribute 9]
	▪ Measurement Method: [Attribute 1–2, 4–8] 1. CPR-01: Ratio of Record with non-blank or non-null field in Product Repository 2. CPR-02: Ratio of NON-Parentless Record in Product Repository (e.g. SKU is referenced by a Product) 3. CPR-03: Ratio of NON-Childless Record in Product Repository (e.g. Product has SKU) 4. TOTAL: 70%xCPR-01 + 15%xCPR-02 +15%xCPR-03 ▪ Frequency: Daily [Attribute 7] ▪ Value: [0–1] [Attribute 5–6] ▪ Expected Threshold:				
ii	Customer could not browse the site conveniently	Customer dissatisfaction	Ambiguous data in Book database (taxonomy mapping problem)	Absence of contradictions (consistency)	Subject, Parent Category
	There are 2 problems here: a. Wrong mapping -> consistency issue b. Different taxonomy -> taxonomy mapping issue For the problem (a) we can use this measurement: ▪ Measurement Method: 1. AOC-01: Ratio of reasonable fields (subject related) Reasonable field: field that has the same top-5 values on the basis of its distribution compared with previous data 2. AOC-02: Ratio of record which adhere business rule example: Title-Subject: Ratio (i) = number of records with Title (i) and Subject (i)/ number of records with Title (i) 3. SC-02: Ratio of record which has non deviated value in Product Repository (List of Values) 4. TOTAL: 15%xAOC-01 + 15%xSC-02 + 70%xAOC-02 ▪ Frequency: Daily ▪ Value: [0–1] ▪ Expected Threshold:				
iii	Unable to run marketing campaign using AdWords and Email channel	Potential revenue loss	Incomplete information in eCommerce system	Completeness per row	All in marketing data model
	▪ see (i) mapping problem				
iv	Internet user could not find the data in top result using search engine	Potential revenue loss	Incomplete information in eCommerce system	Completeness per row	All in websites data model
	▪ see (i) mapping problem				

No	Business Problem	Business Impact	Data Defect	DQ Dimensions	Attribute
v	Offering unavailable product	Customer dissatisfaction, unrecognized revenue, ineffective marketing, and potential revenue loss	a. Inaccurate data in eCommerce system (Journal database)	Accuracy inc. currency	Saleable/ Availability in a Region
			b. Incomplete data in eCommerce system (Journal database)	Completeness, Business Referential Integrity	Fulfillment system
			c. Inconsistent data from Journal database and eCommerce system	Absence of contradiction, Accuracy incl. currency	Product data
			d. Inaccurate data in eCommerce system		Product Data

- Data defect: (a) (Marketing Restriction)
- Measurement Method:
  1. CPR-01: Ratio of Record with non-blank or non-null for availability fields in Product Repository
  2. SC-02: Ratio of record which has non deviated value in Product Repository (List of Values)
  3. AOC-02: Ratio of record which adhere business rule

  rule1:

  Journal

Site	Currency
EST_UK_BS	GBP/ EUR
EST_AU_BS	USD
EST_ASIA_BS	USD
EST_US_BS	USD
EST_JP_BS	JPY/ USD
EST_MEA_BS	USD
EST_EU_BS	EUR

  Book

Site	Currency
EST_UK_BS	GBP/ EUR
EST_AU_BS	AUD/ USD
EST_ASIA_BS	USD
EST_US_BS	USD
EST_JP_BS	JPY/ USD
EST_MEA_BS	USD
EST_EU_BS	EUR

  4. TOTAL = 15%xCPR-01 + 15%xSC-02 + 70%xAOC-02
- Frequency: Daily
- Value: [0–1]
- Expected Threshold:

- Data defect [b]
- Measurement Method:
  1. CPR-01: Ratio of Record with non-blank or non-null field for fulfilment fields in Product Repository
  2. SC-02: Ratio of record which has non deviated value in Product Repository (List of Values)
  3. AOC-02: Ratio of record which adhere business rule

  rule: Journal

Site	Print Journal		eJournal
	PJROMIS	PJARGI	EJSD
EST_AU_BS	DELTA	ARGI	CRM
EST_EU_BS	DELTA	-	CRM
EST_MEA_BS	DELTA	ARGI	CRM
EST_UK_BS	DELTA	-	CRM

No	Business Problem	Business Impact	Data Defect	DQ Dimensions	Attribute

EST_JP_BS	DELTA	-	CRM
EST_US_BS	DELTA	ARGI	CRM
EST_ASIA_BS	DELTA	ARGI	CRM

rule: Book

Site	Print Book	eBook	
	Physical	EBSD	Others
EST_AU_BS	BOOKMASTER	CRM	DELTA
EST_EU_BS	DELTA	CRM	DELTA
EST_MEA_BS	DELTA	CRM	DELTA
EST_UK_BS	DELTA	CRM	DELTA
EST_JP_BS	COPS	CRM	DELTA
EST_US_BS	COPS	CRM	DELTA
EST_ASIA_BS	COPS	CRM	DELTA

4. TOTAL: 15%xCPR-01 + 15%xSC-02 + 70%xAOC-02

- Frequency: Daily
- Value: [0–1]
- Expected Threshold:

- Data defect [c, d]
- Measurement Method:
  1. ACR-01: number of unique ISN in Journal database should be available for eCommerce/ number of unique ISN in eCommerce system -> min/ max, number of unique ISBN in Book database should be available for eCommerce / number of unique ISBN in eCommerce system -> min/ max.
  2. ACR-03: Ratio of Record with exact same value for ISN in Product Repository with data source (Journal database). Ratio of Record with exact same value for ISBN in Product Repository with data source (Book database)
  3. TOTAL: Average of (ACR-01, ACR-03)
- Frequency: Daily
- Value: [0–1]
- Expected Threshold:

vi	Products are not included in the marketing campaign	Potential revenue loss	Taxonomy mapping problem	Absence of contradiction	Subject
	see (ii) mapping problem				

Table 35 Metrics Specification for Preventive and Reactive Measures

No	ID	Measurement Method	Value	Freq.	Attribute
Completeness per row (horizontal completeness) [Attribute 2,10]					
1	CPR-01 [Attribute 1-2, 4-8]	Sebastian-Coleman (Field completeness - non-null able fields), DAMA, Peralta (Semantic Correctness Ratio Metric) **result** = 1 - (Number of row with empty non-null able field divided with number of all row) [Attribute 4]	[0-1] [Attribute 5-6]	Daily [Attribute 7]	▪ String and Numeric ▪ All in websites data model [Attribute 9]
2	CPR-02	Sebastian-Coleman (Parent/child referential integrity)	[0-1]	Daily	▪ String and

No	ID	Measurement Method	Value	Freq.	Attribute
		result = 1 - (Number of unreferenced row (parentless row) divided with number of all rows)			Numeric ▪ All in websites data model
3	CPR-03	Sebastian-Coleman (Child/parent referential integrity) **result** = 1 - (Number of row with empty non-null able reference field and non-exist reference field value (childless row) divided with number of all rows)	[0-1]	Daily	▪ String and Numeric ▪ All in websites data model
4	CPR-04	result = 70%xCPR-01 + 15%xCPR-02 +15%xCPR-03	[0-1]	Daily	
Syntactical correctness (conformity)					
5	SC-01	Sebastian-Coleman (Validity check, single field, detailed results); Peralta (Syntactic Correctness Ratio Metric) **result** = 1 - (Number of row with non-standard value or format divided with number of all rows) ▪ Standard Format: Top-3 string pattern on the basis of distribution OR defined business rule (postcode is 4 char, dash, 2 numeric: ZZZZ-99) ▪ Standard Value: Top-3 value on the basis of distribution OR between min-max value of previous data OR defined business rule (price is >=0)	[0-1]	Daily	▪ Numeric: between min-max value ▪ String and Numeric: business rule, string patter, top-3 value
6	SC-02	Sebastian-Coleman (Validity check, single field, detailed results:), Peralta (Syntactic Correctness Deviation Metric) **result** = 1 - (Number of row with deviated value divided with number of all rows) ▪ Non-deviated value: there is a similar value at reference table with similarity>=0.8 for example (Levenshtein distance/length of longer string) <=0.2 OR Jaro-Winkler distance>=0.8. ▪ Similarity=1 for numeric type field	[0-1]	Daily	String and Numeric (deviation=0)
7	SC-03	**result** = Average (SC-01, SC-02) if the reference table for SC-02 is not available then SC-03 = SC-01	[0-1]	Daily	String and Numeric (deviation=0)
	NOTE: ▪ SC-01: Non LoV, Incorrect values include: non empty value that could be considered as blank, e.g., Text:"UNKNOWN", Text:"EMPTY", Date: "1/1/1900 00:00:00", Numeric:"0" ? ▪ Reference Table (LoV) in PIM for SC-02: Business Classification, Country, Imprint, Language, Legal Entity, Page Count Type, Product Distribution Type, Product Manifestation Type, Product Type, Publisher, Region, State, Subject Area, Subject Area Type				
Absence of contradictions (consistency) and normative consistency					
8	AOC-01	Sebastian-Coleman (Consistent column profile) **result** = 1 - (number of non-reasonable fields divided with number of all fields) ▪ Reasonable field: field that has the same top-5 values on the basis of its distribution compared with previous data	[0-1]	Daily	String
9	AOC-02	Sebastian-Coleman (Consistent dataset content, distinct count of represented entity, with ratios to record counts;	[0-1]	Daily	String and Numeric

No	ID	Measurement Method	Value	Freq.	Attribute
		Consistent cross table multi columns profile:) ▪ Rules: Title - Category, Price - Location, Location/ Type - Fulfillment Company Code **result =** average (all ratio per avail)			
10	AOC-03	Sebastian-Coleman (Consistent record counts by aggregated date) ▪ val1: (M-1 rows/M-2 rows); val2: (last year M-1 rows/ M-2 rows) ▪ val3 = val1/ val2 ▪ minVal = min(val1,val2,val3); maxVal=max(val1,val2,val3) ▪ rawVal = not (minVal or maxVal) ▪ **result =** (rawVal-minVal) / (maxVal-minVal) Quarterly: change M with Q	[0-1]	Monthly or Quarterly	String and Numeric
11	AOC-04	**result =** Average (SC-02, AOC-01, AOC-02, AOC-3)	[0-1]	Daily	String and Numeric
Absence of repetitions (free of duplicates)					
12	AOR-01	**result =** 1 - (Number of duplicate row divided with number of all unique rows) Unique = unique ISBN for book or unique ISSN for journal	[0-1]	Daily	String and Numeric
Business referential integrity (integrity)					
13	BRI-01	**result =** Average (CPR-02, CPR-03, AOC-02)	[0-1]	Daily	String and Numeric
	NOTE: This measurement could be ignored since it is composed from other measurement's components.				
Accuracy incl. currency					
14	ACR-01	Peralta (Semantic Correctness Ratio Metric), DAMA. **Result** = average(number of unique book ISBN in Book database + number of unique book ISBN in French site XML/ number of unique book ISBN in eCommerce system, number of unique ISSN in Journal database/ number of unique ISSN in eCommerce system)	[0-1]	Daily	String and Numeric
15	ACR-02	DAMA, Sebastian-Coleman (Timely delivery of data for processing, Timely availability of data for access) ▪ Ratio 1: 1- (min difference of time data in Journal database/Book database/French XML with time data in eCommerce system)/ 720 ▪ Ratio 2: 1- (min difference of time data in eCommerce system and time data in Web)/ 720 ▪ **result** = average(Ratio 1, Ratio 2) ▪ 720 minutes = 12 hours, if difference>720 then Ratio (i) = 0	[0-1]	Daily	String and Numeric
16	ACR-04	**Result** = average(CPR-01, SC-01, SC-02, ACR-01, ACR-02)	[0-1]	Daily	String and Numeric
	NOTE: ▪ ACR-01: Book database, Journal database, and French Site XML is considered as the "Real World." In MDM, the repository holds the golden record. There could be other "Real Word" entities if the architectural type is Transaction Hub.				

Appendix 12 Data Quality Metrics Assessment on eCommerce Database for Phase III

Appendix 13 Phase IV: Data Quality Metrics Integration

References

[1] Batini, Carlo, and Monica Scannapieca. Data quality: concepts, methodologies and techniques. Springer, 2006.

[2] Batini, Carlo et al. "A Framework And A Methodology For Data Quality Assessment And Monitoring." ICIQ. 2007.

[3] Batini, Carlo et al. "Methodologies for data quality assessment and improvement." ACM Computing Surveys (CSUR) 41.3 (2009): 16.

[4] Batini, Carlo, Maurizio Lenzerini, and Shamkant B. Navathe. "A comparative analysis of methodologies for database schema integration." ACM computing surveys (CSUR) 18.4 (1986): 323-364.

[5] DAMA. (2009). The DAMA guide to the data management body of knowledge. Bradley Beach, NJ: Technics Publications.

[6] Dreibelbis, Allen et al. Enterprise master data management: an SOA approach to managing core information. Pearson Education, 2008.

[7] English, Larry P. Improving data warehouse and business information quality: methods for reducing costs and increasing profits. Vol. 1. New York: Wiley, 1999

[8] Flanagin, Andrew J. et al. "User-generated ratings and the evaluation of credibility and product quality in ecommerce transactions." System Sciences (HICSS), 2011 44th Hawaii International Conference on. IEEE, 2011.

[9] Flyvbjerg, Bent. "Five misunderstandings about case-study research." Qualitative inquiry 12.2 (2006): 219-245.

[10] Haug, Anders et al. "Master data quality barriers: an empirical investigation." Industrial Management & Data Systems 113.2 (2013): 234-249.

[11] Heinrich, Bernd, Marcus Kaiser, and Mathias Klier. "How to measure data quality? A metric-based approach." (2007).

[12] Hüner, Kai M. et al. "Methode zur Spezifikation geschäftsorientierter Datenqualitätskennzahlen." Institut für Wirtschaftsinformatik, Universität St. Gallen, St. Gallen (2011).

[13] Lee, Yang W. et al. "AIMQ: a methodology for information quality assessment." Information & management 40.2 (2002): 133-146.

[14] Loshin, David. Master data management. Morgan Kaufmann, 2010.

[15] Loshin, David. The practitioner's guide to data quality improvement. Access Online via Elsevier, 2010.

[16] Molla, Alemayehu, and Paul S. Licker. "E-Commerce Systems Success: An Attempt to Extend and Respecify the Delone and MaClean Model of IS Success." J. Electron. Commerce Res. 2.4 (2001): 131-141.

[17] Morbey, Guilherme. Data Quality for Decision Makers: A dialog between a board member and a DQ expert 2nd edition. Springer Gabler, 2013.

[18] Otto, Boris. "How to design the master data architecture: Findings from a case study at Bosch." International Journal of Information Management 32.4 (2012): 337-346.

[19] Otto, Boris, Kai M. Hüner, and Hubert Österle. “Identification of Business Oriented Data Quality Metrics.” ICIQ. 2009.

[20] Perlman, Yael. “Causal Relationships in the Balanced Scorecard: A Path Analysis Approach.” Journal of Management & Strategy 4.1 (2013).

[21] Peralta, Verónika. “Data freshness and data accuracy: A state of the art.” Instituto de Computacion, Facultad de Ingenieria, Universidad de la Republica, Uruguay, Tech. Rep. TR0613 (2006).

[22] Pipino, Leo L., Yang W. Lee, and Richard Y. Wang. “Data quality assessment.” Communications of the ACM 45.4 (2002): 211-218.

[23] Sebastian-Coleman, Laura. Measuring Data Quality for Ongoing Improvement. Morgan Kaufmann. 2013

[24] Silvola, Risto et al. “Managing one master data–challenges and preconditions.” Industrial Management & Data Systems 111.1 (2011): 146-162.

[25] Reeve, April. Managing Data in Motion: Data Integration Best Practice Techniques and Technologies. Access Online via Elsevier, 2013.

[26] Tsai, Yuan-Cheng, and Yu-Tien Cheng. “Analyzing key performance indicators (KPIs) for E-commerce and Internet marketing of elderly products: A review.” Archives of gerontology and geriatrics 55.1 (2012): 126-132.

[27] Verschuren, Piet, Hans Doorewaard, and M. J. Mellion. Designing a research project. Eleven International Publishing, 2010.

[28] Wang, Richard Y., Diane M. Strong, and Lisa Marie Guarascio. “Beyond accuracy: What data quality means to data consumers.” J. of Management Information Systems 12.4 (1996): 5-33.

[29] Wang, Richard Y., Martin P. Reddy, and Henry B. Kon. “Toward quality data: An attribute-based approach.” Decision Support Systems 13.3 (1995): 349-372.

[30] Woodall, Philip, Alexander Borek, and Ajith Kumar Parlikad. “Data Quality Assessment: The Hybrid Approach.” Information & Management (2013).