Phase IV: DQ Metrics Integration for Product MDM

REFERENCE DOCUMENT

by
Marhendra Lidiansa
l.marhendra@student.tudelft.nl

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 4

List of Figures 5
Document Control 6
1 Overview 7
2 Phase I Integration 9
2.1 Data Structure Mapping and Conformance 9
2.1.1 Pre Integration 9
2.1.2 Schema Comparison 9
2.1.3 Conforming, Merging and Structuring the Schema 11
2.2 Data Quality Metrics Integration 13
2.2.1 Activities 13
2.2.2 Result of Metrics Integration 14
3 Phase II Integration 20
3.1 Data Structure Mapping and Conformance 20
3.1.1 Pre Integration 20
3.1.2 Schema Comparison 20
3.1.3 Conforming, Merging and Structuring the Schema 23
3.2 Data Quality Metrics Integration 26
3.2.1 Activities 26
3.1 Result 27
3.1.1 Completeness 27
3.1.2 Syntactical Correctness 28
3.1.3 Absence of Contradiction 31
3.1.4 Absence of Repetition 33
3.1.5 Accuracy incl. Currency 33
Appendix 1. Data Quality Metrics Specification for e-commerce System 35
Appendix 2. Data Quality Metrics Specification for Customer System 41
Appendix 3. Discrepancy Assessment Results for SC-01 44
Appendix 4. Discrepancy Assessment Results for SC-02 46
Appendix 5. Data Analytics to Select Important Attributes 51
Appendix 6. Workshops Documents 54
References 55

List of Tables

Table 1 Mapped Entities for e-commerce and Customer System Book 9
Table 2 Mapped Entities for e-commerce and Customer System Journal 9
Table 3 Unmapped Attributes for e-commerce and Customer System 10
Table 4 Entities - Tables Mapping 11
Table 5 Conflict Resolution for Phase I Schema Integration 12
Table 6 Metrics Availablity 13
Table 7 Developed DQ Metrics Attributes 14
Table 8 Journal e-commerce Attributes for SC-01 15
Table 9 Book e-commerce Attributes for SC-01 15
Table 10 Journal CS Attributes for SC-01 15
Table 11 Book CS Attributes for SC-01 16
Table 12 Integrated SC-01 for Journal (e-commerce and Customer System) 17
Table 13 Integrated SC-01 for Books (e-commerce and Customer System) 17
Table 14 SC-02 for Journal (e-commerce System and CS) 18
Table 15 SC-02 for Book (e-commerce System and CS) 19
Table 16 Mapped Attributes for Integrated Applications and PIM 20
Table 17 Unmapped Attributes for Integrated Applications and PIM 22
Table 18 Entities - Tables Mapping 23
Table 19 Conflict Resolution for Phase II Schema Integration 24
Table 20 Add Attributes to PIM 26
Table 21 Metrics Availability Phase II 26
Table 22 SC-01 for Journal (PIM) 28
Table 23 SC-01 for Books (PIM) 29
Table 24 SC-02 for Book and Journal (PIM) 30
Table 25 AOC-01 for PIM 32
Table 26 AOC-02 for PIM 33
Table 27 Metrics Specification for Business Problems 35
Table 28 Metrics Specification for Preventive and Reactive Measures 38
29 Data Quality Metrics for Customer System 41
Table 30 SC-01 Rules for Journal 42
Table 31 SC-01 Rules for Book 42
Table 32 e-commerce-Customer System Journal Data Discrepancies for SC-01 44
Table 33 e-commerce-Customer System Book Data Discrepancies for SC-01 44
Table 34 Journal Imprint Discrepancies 46
Table 35 Book Product Type Discrepancies 48
Table 36 Book Imprint Discrepancies 48
Table 37. Book Info CPR-01 Assessment 51
Table 38 Book Regional Info CPR-01 Assessment 51
Table 39 Confusion Matrix 52
Table 40 Data Analytics Result 52

List of Figures

Figure 1 Steps of identifying quality attributes, Wang, et al. [3] 7
Figure 2 Data Model with Quality (Attribute Level), Wang, et al. [3] 7
Figure 3 Simplified ERD for e-commerce 11
Figure 4 PIM Entities and Attributes 23

Document Control

Document Version : 1.0

Version history

Version	Final/Draft	Release date	Author	Comments
1.0	Final	21-May-2014	Marhendra L	• This document is a reference document for the Main Report • This Document describes the phase IV: Data Qualty Metrics Integration

1 Overview

Master data objects are distributed across numerous enterprise applications where each application has its own data models and data life cycle process. Thus, data quality specification in MDM should meet the requirements of several business applications (Loshin [2]). Wang, et al. [3] developed the process model to build the integrated data schema with the quality definition from several application views as described in Figure 1. The output of having activity no 1 until 3 is similar to the general process framework, namely the list of data quality metrics. Thus, this thesis uses the 4th activity to develop the data quality metrics for product MDM by integrating data quality metrics of two systems.

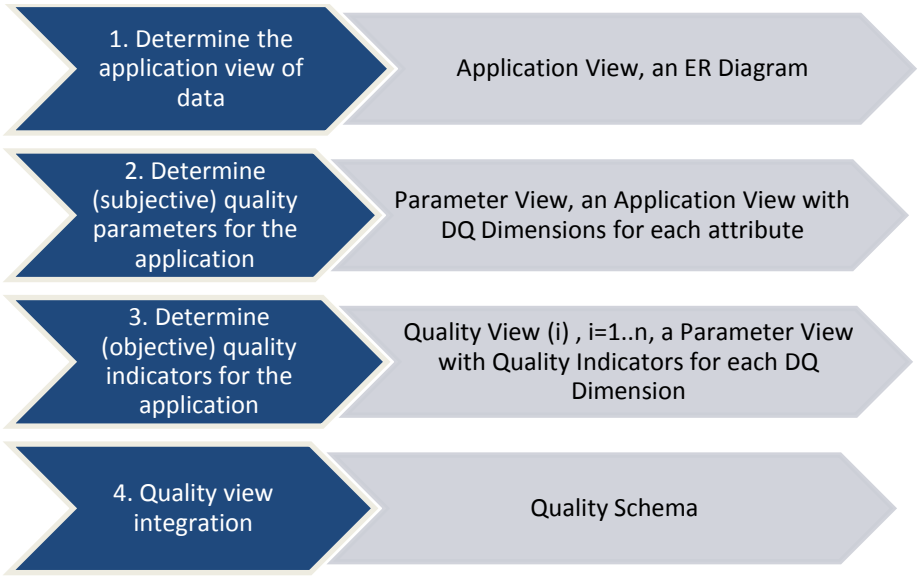

Figure 1 Steps of identifying quality attributes, Wang, et al. [3]

In order to apply the process model, the thesis needs to develop the data model that represents the model in Figure 2. This data model, with the cell level tagging, is argued by Wang, et al. [3] to meet the needs of multidimensionality and hierarchical of data quality. They put the tagging at the cell level because 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.

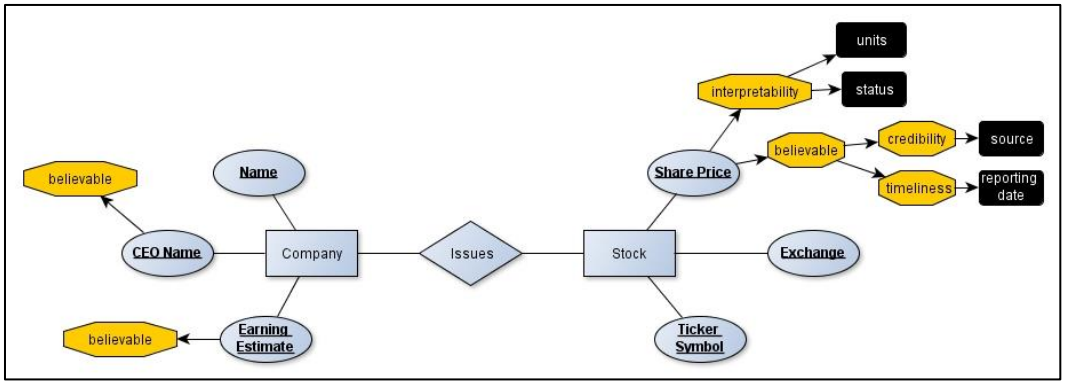

Figure 2 Data Model with Quality (Attribute Level), Wang, et al. [3]

This thesis work has 2 cases, namely e-commerce System that is mainly used in the validation phase, and 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. The data quality metrics specification for e-commerce System is in Appendix 1 and the specification for Customer System is in Appendix 2. In order to conduct the 4th activity, Wang, et al. [3] used the process model developed by Batini, et al. [1] to integrate the schemas as follows:

1. Pre Integration
   An analysis of schemas is carried out to provide the information needed to make these decisions: 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 these qualitative criteria:
   a. Completeness and Correctness
      The integrated schema must contain all concepts present in any component schema correctly.
   b. Minimality
      A concept must be represented once in the integrated schema if it is represented in more than one component schema.
   c. Understandability
      Among several possible results, the one that is (qualitatively) the most understandable should be chosen.

The integration process involves two components, namely the data schema and the data quality metrics. Elsevier also has developed a product MDM data model from several application views, but the model is without data quality information. Thus, the process is split into 2 phases as follows:

1. Phase 1. Integration of Customer System and e-commerce DQ Metrics (Application Views)
2. Phase 2. Integration of Application Views and Product MDM.

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, and
2. The product data model for the MDM is developed on the basis of the requirement of n (>2) other applications.

2 Phase I Integration

The phase I Integration goal is to integrate the data quality metrics of e-commerce System and Customer System. This thesis follows the activities described by Batini, et al. [1], namely Pre Integration, Comparison of the Schema, Conforming the Schema, and Merging and Structuring. Because the data quality metrics has two main components, namely the data model and metrics specification, this thesis also separates the activities for each component.

2.1 Data Structure Mapping and Conformance

2.1.1 Pre Integration

The pre integration activity is using the information from the data quality metrics development for the 2 study cases. Each study case provides a list of entities and attributes for the books and journal.

2.1.2 Schema Comparison

This activity compares the database schema of e-commerce and Customer System and provides two results, namely the mapped entities in Table 1 and Table 2, and unmapped entities in Table 3.

Table 1 Mapped Entities for e-commerce and Customer System Book

Book	e-commerce System			Customer System			
	Column Name	DATA TYPE	SIZE	Column Name	DATA TYPE	SIZE	Req.
	DCSX_PRODUCT						
Title	TITLE	VARCHAR2	256	Title	TEXT	255	Y
Product Type	PRODUCT_TYPE	VARCHAR2	40	ContentType	TEXT	255	Y
Print Book/ eBook ISBN	ISBN	VARCHAR2	20	FormattedISBN	TEXT	255	Y
Author Name	ALL_AUTHOR	VARCHAR2	4000	Author/ Editor	TEXT	255	N
	ELS_PRODUCT						
Imprint	IMPRINT	VARCHAR2	50	ImprintPublisher	TEXT	255	Y
Edition	EDITION_NUMBER	VARCHAR2	80	EditionText	TEXT	255	N
	ELS_PRODUCT_EN						
SubTitle	SUB_TITLE	VARCHAR2	255	SubTitle	TEXT	255	N
	ELS_PRODUCT_ REGIONAL_INFO						
Publication Date	PUB_DATE	DATE		Publication Year	TEXT	255	Y

*Req. = Required

Table 2 Mapped Entities for e-commerce and Customer System Journal

	e-commerce System			Customer System			
Journal	Column Name	DATA TYPE	SIZE	Column Name	DATA TYPE	SIZE	Req.
	DCSX_PRODUCT						
Title	TITLE	VARCHAR2	256	Name	TEXT	255	Y
Product Type	PRODUCT_TYPE	VARCHAR2	40	ContentType	TEXT	255	Y
ISSN	ISBN	VARCHAR2	20	ISSN	TEXT	255	Y

	e-commerce System			Customer System			
Journal	**Column Name**	**DATA TYPE**	**SIZE**	**Column Name**	**DATA TYPE**	**SIZE**	**Req.**
	ELS_PRODUCT						
Issue Number	FREQUENCY	VARCHAR2	100	IssuesPerYear	DOUBLE		Y

*Req. = Required

Table 3 Unmapped Attributes for e-commerce and Customer System

Book	Journal	COLUMN_NAME	DATA TYPE	SIZE	Req
E-COMMERCE SYSTEM					
		ELS_PRODUCT_EN			
	Subtitle	SUBTITLE	VARCHAR	255	Y
		ANG_PRODUCT			
	Impact Factor	IMPACT_FACTOR	NUMBER (10,3)	7	Y
		DCSX_PRODUCT			
Number of Page		NUMBER_OF_PAGES	VARCHAR2	50	Y
Dimensions		PUB_NUM_LOG	VARCHAR2	16	Y
	5yr Impact Factor	VERSION_NUMBER	VARCHAR2	40	Y
Table of Contents		TABLE_OF_CONTENTS	CLOB	~	Y
		ELS_PRODUCT			
Next Edition		NEXT_EDITION_ISBN	VARCHAR2	20	Y
Authors-Name	Editor Information	AUTHOR_ALIST	CLOB	~	Y
	Editorial Information	AUTHOR_BLIST	CLOB	~	Y
	Valume Number	VOLUME_NUMBER	VARCHAR2	20	Y
		ELS_PRODUCT_EN			
Price-Format	Price-Format	WEB_PRODUCT_ TYPE_NAME	VARCHAR2	100	Y
	Abstract and Indexing	ABSTRACT	CLOB	~	Y
		DCSX_SKU			Y
Price-Format	Price-Format	SKU_TYPE	VARCHAR2	20	Y
		ELS_DCSX_SKU			
		FULFILLMENT_ COMPANY_CODE	VARCHAR2	40	Y
		ELS_SKU			
Stock	Stock	OUT_OF_PRINT	NUMBER (1)	1	Y
Price-Format	Price-Format	PURCHASE_TYPE	VARCHAR2	75	Y
		DCS_PRICE			
Price	Price	PRICE_LIST, DISPLAY_NAME	VARCHAR2	40, 254	Y
Price	Price	LIST_PRICE	NUMBER (19,7)	12	Y
		ELS_PRODUCT_ REGIONAL_INFO			
Stock	Stock	PUB_STATUS	INTEGER	38	Y
Short Description	Short Description	SHORT_DESCRIPTION	CLOB	~	Y
Key Feature		KEY_FEATURE	VARCHAR2	4000	Y

Book	Journal	COLUMN_NAME	DATA TYPE	SIZE	Req
Long Description	Aim and Scope	LONG_DESCRIPTION	CLOB	~	Y
Readership	Audience	AUDIENCE	CLOB	~	Y
Review		QUOTES	CLOB	~	Y
		FULFILLMENT_ COMPANY_CODE	VARCHAR2	40	Y
[drop down site]	[drop down site]	SITE_ID	VARCHAR2	40	Y
		DCS_PRD_PRNT_CATS			
Breadcrumb		CATEGORY_ID	VARCHAR2	40	Y
Breadcrumb		DISPLAY_NAME	VARCHAR2	254	Y
CUSTOMER SYSTEM					
CopyrightText	CopyrightText	CopyrightText	TEXT	255	Y
SmallCover	SmallCover	SmallCover	TEXT	255	N
HomePageURL	HomePageURL	HomePageURL	TEXT	255	N
	VolumeIssueAlert Available	VolumeIssueAlert Available	TEXT	255	Y
	Submityour ArticleURL	Submityour ArticleURL	TEXT	255	Y
	SocietyURL	SocietyURL	TEXT	255	N
	SocietyText	SocietyText	TEXT	255	Y/N

*Req. = Required

2.1.3 Conforming, Merging and Structuring the Schema

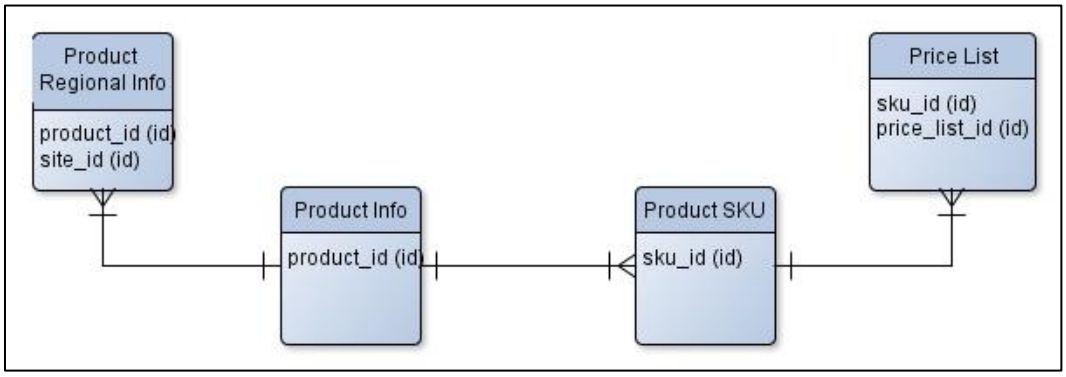

Figure 3 Simplified ERD for e-commerce

The e-commerce has generic data model as described in Figure 3 and we could map the tables for each case to the entities in Table 4. This mapping is used a guidance to conform the schema where the attributes from Customer System are for Product Info entity.

Table 4 Entities - Tables Mapping

Entities	e-commerce Tables	Customer System Table
Product Info	DCSX_PRODUCT, ANG_PRODUCT, ELS_PRODUCT, ELS_PRODUCT_EN, DCS_PRD_PRNT_CAT, and DCS_CATEGORY	CSRT (JOURNAL) and NONSERIALSRT (BOOK)
Product Regional	ELS_PRODUCT_REGIONAL_INFO	NA

Entities	e-commerce Tables	Customer System Table
Info		
Product SKU	DCSX_SKU, ELS_SKU, and ELS_DCSX_SKU	NA
Pricing List	DCS_PRICE, and DCS_PRICE_LIST	NA

The unmapped attributes in Table 3 are all used in the integrated schema because each attribute has specific data quality metrics. Thus, the end schema will have the generic model in Figure 3. Within the mapped attributes in Table 2, there are several discrepancies that need to be resolved using the qualitative criteria as follows:

1. 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 criteria. When comparing the column name, this thesis work selects the one that describes the content better.
2. 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 value compound process could be conducted at the application level or data exchange.
3. 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.

Table 5 Conflict Resolution for Phase I Schema Integration

WEB		Conflict Source	Resolution				Description
Book	**Journal**		**Column Name**	**DATA TYPE**	**SIZE**	**Req**	
			DCSX_PRODUCT				
Title	Title	Data Size	TITLE	VARCHAR2	256	Y	▪ Data Size = ES
Product Type	Product Type	Column Name and Data Size	PRODUCT_TYPE	VARCHAR2	255	Y	▪ Column Name = ES ▪ Data Size = CS
Print Book/ eBook ISBN	ISSN	Column Name and Data Size	ISBN/ ISSN	VARCHAR2	255	Y	▪ Column Name = ES ▪ Data Size = CS
Author Name		Column Name, Data Size, and Required	AUTHOR NAME	VARCHAR2	4000	Y	▪ Column Name = CS ▪ Data Size = ES
			ELS_PRODUCT				
Imprint		Column Name and Data Size	IMPRINT	VARCHAR2	255	Y	▪ Column Name = ES ▪ Data Size = CS
Edition		Column Name and Data Size	EDITION_NUMBER	VARCHAR2	255	Y	▪ Column Name = ES ▪ Data Size = CS
	Issue Number	Column Name, Data Size, and Required	ISSUESPERYEAR	DOUBLE		Y	▪ Column Name = CS ▪ Data Size =CS
			ELS_PRODUCT_EN				

WEB		Conflict Source	Resolution				Description
Book	Journal		Column Name	DATA TYPE	SIZE	Req	
SubTitle	SubTitle	Column Name and Data Size	SUBTITLE	VARCHAR2	255	Y	▪ Column Name = CS ▪ Data Size = CS
			ELS_PRODUCT_ REGIONAL_INFO				
Publication Date		Column Name, Data Type, and Size	PUB_DATE	DATE	DATE	Y	▪ Column Name = ES ▪ Data Type = ES

*ES = e-commerce System , CS = Customer System

2.2 Data Quality Metrics Integration

2.2.1 Activities

This activity integrates the metrics specification for the new integrated schema. Because each attributeis attached to certain data quality metrics, we will describe the integration process from the metrics perspective. This process also uses the activities for data structure integration as follows:

2.2.1.1 Pre Integration

The pre integration activity is using the information from the data quality metrics development for the 2 study cases. Each study case provides a list of data quality metrics for the books and journal data.

2.2.1.2 Schema Comparison

This activity compares the metrics from both applications to find unmapped and mapped metrics as in Table 6.

Table 6 Metrics Availablity

Metrics	e-commerce System	Customer System
Completeness per Row		
CPR-01	Yes	Yes
CPR-02	Yes	NO
CPR-03	Yes	NO
Syntactical Correctness		
SC-01	Yes	Yes
SC-02	Yes	Yes
Absence of Contradiction		
AOC-01	Yes	NO
AOC-02	Yes	NO
Absence of Repetition		
AOR-01	Yes	Yes
Accuracy Inc. Currency		
ACR-01	Yes	NO
ACR-02	Yes	NO

2.2.1.3 Conforming, Merging, and Structuring the Schema

Within the mapped metrics in Table 6, there are several discrepancies that need to be resolved. The resolution change several attributes of DQ metrics in Table 7, namely data and rules in measurement methods. The resolution for the data quality metrics also uses the qualitative criteria as follows:

1. Different List of Attributes
   Using the completeness, we need to combine the attributes from both applications.
2. Different Data Rules
   This thesis uses the completeness criteria to resolve the conflict because the rules itself are not conflicting. The data rules are used in syntactical correctness data quality metrics.

Table 7 Developed DQ Metrics Attributes

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

2.2.2 Result of Metrics Integration

2.2.2.1 Completeness Per Row

a. CPR-01
   Updated DQ Attributes: Data
   - Table 5 : All attributes in Resolution
   - Table 3 :
     - All in e-commerce
     - All in Customer System with Req. = 'Y.' (Product Info entity)

b. CPR-02
   Use e-commerce System because it is not applicable for Customer System.

c. CPR-03
   Use e-commerce System because it is not applicable for Customer System.

2.2.2.2 Syntactical Correctness

a. SC-01

Updated DQ Attributes: Data and Rules for Measurement Method

i. Data and Rules for SC-01 in e-commerce System

- Rules
  - Text : not in (“UNKNOWN”, “EMPTY”, “BLANK”, “NULL”)
  - Numeric : >0
  - Date : not (1 Jan 1900 or 1 Jan 1970)
- Attributes :

Table 8 Journal e-commerce Attributes for SC-01

Attributes	Type
Info	
TITLE	Text
ISBN	Text
VERSION_ NUMBER	Numeric
AUTHOR_ALIST	Text
AUTHOR_BLIST	Text
VOLUME_ NUMBER	Text
EDITOR	Text
FREQUENCY	Text
ABSTRACT	Text
SUB_TITLE	Text
IMPACT_FACTOR	Numeric
Regional Info	
SHORT_ DESCRIPTION	Text
AUDIENCE	Text
Price	
LIST_PRICE	Numeric

Table 9 Book e-commerce Attributes for SC-01

Attributes	**Type**
Info	
TITLE	Text
ISBN	Text
ALL_AUTHOR	Text
NUMBER_OF_ PAGES	Text
PUB_NUM_ LOG	Text
TABLE_OF_ CONTENTS	Text
IMPRINT	Text
NEXT_EDITION_ ISBN	Text
AUTHOR_ALIST	Text
SUB_TITLE	Text
Regional Info	
PUB_DATE	Date
SHORT_ DESCRIPTION	Text
KEY_FEATURE	Text
LONG_ DESCRIPTION	Text
AUDIENCE	Text
QUOTES	Text
Price	
LIST_PRICE	Numeric

ii. Data and Rules for SC-01 in Customer System

- Journal

Table 10 Journal CS Attributes for SC-01

No	Field	Req.	Description (Rules for Correctness)
1	Name	Y	All accented characters must be coded as Unicode Hexadecimal codes. E.g. the **á** should be encoded as **á** in the sheet; the **&** as

No	Field	Req.	Description (Rules for Correctness)
			& Please captitalize the first character of each noun and adjective, and of the very first word on the title. E.g. “the title of the new journal” becomes “The Title of the New Journal”
2	Issues per year	Y	>0
3	Small Cover	Y	Must always be “S<issn>.gif, where <issn> is the unformatted ISSN – i.e. without hyphen – of the publication in question
4	ISSN	Y	Copy ISSN from IGT form, but remove dash (!). E.g. 12345678 and not 1234-5678.

- Book

Table 11 Book CS Attributes for SC-01

No	Field	Req.	Description (Rules for Correctness)
1	Title	Y	The Title of the book without any reference to edition number/text or subtitle. Glyphs that are non-standard (e.g. & and ©) or accented letters (e.g. Ü or ê) MUST be expressed in Unicode with this format : **&#x**<Unicode_value> For example: & should be expressed as & (or &) © as © (or ©) Ü as ü (or ü) ê as ê (or ê) See **Error! Reference source not found.**: Unicode Charts for the most common ones.
2	Subtitle	N	Glyphs that are non-standard (e.g. & and ©) or accented letters (e.g. Ü or ê) MUST be expressed in Unicode with this format : **&#x**<Unicode_value> See **Error! Reference source not found.**: Unicode Charts for the most common ones.
3	Formatted ISBN	Y	17 character with 4 hypen and 13 number : XXX-X-XX-XXXXXX-X
4	Small Cover	N	The SC value is made up of the unformatted ISBN with the suffix “_cov150h” and the file extension .gif. For example, 9780080444109_cov150h.gif

iii. Integrated Data and Rules for SC-01

- Journal

Table 12 Integrated SC-01 for Journal (e-commerce and Customer System)

Attributes	Rules
Info	
TITLE	a. Text in e-commerce System b. All accented characters must be coded as Unicode Hexadecimal codes. E.g. the á should be encoded as á in the sheet; the & as & c. Please captitalize the first character of each noun and adjective, and of the very first word on the title. E.g. “the title of the new journal” becomes “The Title of the New Journal”
ISSN	a. Text in e-commerce System b. Copy ISSN from IGT form, but remove dash (!). E.g. 12345678 and not 1234-5678.
VERSION_ NUMBER	Numeric in e-commerce System
AUTHOR_ALIST	Text in e-commerce System
AUTHOR_BLIST	Text in e-commerce System
VOLUME_ NUMBER	Text in e-commerce System
EDITOR	Text in e-commerce System
ISSUES_PER_YEAR	Numeric in e-commerce System (>0)
ABSTRACT	Text in e-commerce System
SUB_TITLE	Text in e-commerce System
IMPACT_FACTOR	Numeric in e-commerce System
Regional Info	
SHORT_ DESCRIPTION	Text in e-commerce System
AUDIENCE	Text in e-commerce System
Price	
LIST_PRICE	Numeric in e-commerce System

- Book

Table 13 Integrated SC-01 for Books (e-commerce and Customer System)

Attributes	**Type**
Info	
TITLE	• Text in e-commerce System • The Title of the book without any reference to edition number/text or subtitle. • Glyphs that are non-standard (e.g. & and ©) or accented letters (e.g. Ü or ê) MUST be expressed in Unicode with this format : **&#x**<Unicode_value> • For example: & should be expressed as & (or &) © as © (or ©) Ü as ü (or ü) ê as ê (or ê)
ISBN	Text in e-commerce System
AUTHOR	Text in e-commerce System
NUMBER_OF_ PAGES	Text in e-commerce System
PUB_NUM_ LOG	Text in e-commerce System
TABLE_OF_ CONTENTS	Text in e-commerce System

Attributes	**Type**
IMPRINT	Text in e-commerce System
NEXT_EDITION_ ISBN	Text in e-commerce System
AUTHOR_ALIST	Text in e-commerce System
SUB_TITLE	• Text in e-commerce System • Glyphs that are non-standard (e.g. & and ©) or accented letters (e.g. Ü or ê) MUST be expressed in Unicode with this format : &#x<Unicode_value>
Regional Info	
PUB_DATE	Date in e-commerce System
SHORT_ DESCRIPTION	Text in e-commerce System
KEY_FEATURE	Text in e-commerce System
LONG_ DESCRIPTION	Text in e-commerce System
AUDIENCE	Text in e-commerce System
QUOTES	Text in e-commerce System
Price	
LIST_PRICE	Numeric in e-commerce System

b. SC-02

Updated DQ Attributes: Data and List of Values for Measurement Method

i. SC-02 mapping for e-commerce System and CS

- Journal

Table 14 SC-02 for Journal (e-commerce System and CS)

Attribute	**e-commerce System**	**Customer System**
Info		
PRODUCT_TYPE	v	v
WEB_PRODUCT_ TYPE_NAME	v	-
CATEGORY_ID	v	-
Regional Info		
PUB_STATUS	v	-
FULFILLMENT _COMPANY _CODE	v	-
SITE_ID	v	-
SKU		
SKU_TYPE	v	-
FULFILLMENT _COMPANY _CODE	v	-
OUT_OF_PRINT	v	-
PURCHASE_TYPE	v	-
Price		
PRICE_LIST	v	-
Imprint	-	v
Sample Issue Available	-	v
Journal/Book Alert Available	-	v

- Book

Table 15 SC-02 for Book (e-commerce System and CS)

Attribute	e-commerce System	CS
Info		
PRODUCT_TYPE	v	v
IMPRINT	v	v
WEB_PRODUCT _TYPE_NAME	v	-
CATEGORY_ID	v	-
Regional Info		
PUB_STATUS	v	-
FULFILLMENT _COMPANY_CODE	v	-
SITE_ID	v	-
SKU		
SKU_TYPE	v	-
FULFILLMENT _COMPANY_CODE	v	-
OUT_OF_PRINT	v	-
PURCHASE_TYPE	v	-
Price		
PRICE_LIST	v	-

2.2.2.3 Absence of Contradiction

Use e-commerce because it is not applicable for Customer System

2.2.2.4 Absence of Repetition

Both system use the same method.

2.2.2.5 Accuracy incl. Currency

Use e-commerce because it is not applicable for Customer System.

3 Phase II Integration

The phase II Integration goal is to integrate the data quality metrics of product MDM data model (PIM) and integrated applications in Phase I. This process follows the same activities used in previous phase.

3.1 Data Structure Mapping and Conformance

3.1.1 Pre Integration

The pre integration activity is using the information from the data quality metrics development for the 2 study cases. Each study case provides a list of entities and attributes for the books and journal.

3.1.2 Schema Comparison

This activity compares the database schema of product MDM and integrated applications in Phase II. and provides two information, namely the mapped entities in Table 16 and unmapped entities in Table 17.

Table 16 Mapped Attributes for Integrated Applications and PIM

Web		e-commerce and Customer System			PIM			
Book	**Journal**	**Attibute**	**Data Type**	**Data Size**	**PIM Entity**	**PIM Attribute**	**Data Type**	**Data Size**
		DCSX_PRODUCT						
Title	Title	TITLE	VARCHAR2	256	Product	Product Title	VARCHAR	400
Product Type	Product Type	PRODUCT_ TYPE	VARCHAR2	255	Product, Product Type	Product Type Code, Product Type Name	VARCHAR	35, 400
Print Book/ eBook ISBN	ISSN	ISBN	VARCHAR2	255	Product	ISBN, ISSN	VARCHAR	400
Author Name		AUTHOR	VARCHAR2	4000	Books	Primary Author Name	VARCHAR	400
Number of Page		NUMBER_ OF_PAGES	VARCHAR2	50	Page Count	Page Count Page Quantity	INTEGER	38
Dimensions		PUB_NUM_LOG	VARCHAR2	16	Print Product	Page Height Amount, Page Width Amount	NUMERIC (15,2)	13
		ELS_PRODUCT						
Imprint		IMPRINT	VARCHAR2	255	Product, Imprint	Imprint Code, Imprint Name	VARCHAR	35, 400
Next Edition		NEXT_ EDITION_ ISBN	VARCHAR2	20	Product Link	Product Link Type Code, Related Product Id		
Authors-Name	Editor Information	AUTHOR_ALIST	CLOB	~	Party in Product, Party	Person Display Name	VARCHAR	400
	Editorial Information	AUTHOR_BLIST	CLOB	~	Party in Product, Party	Person Display Name	VARCHAR	400
	Valume Number	VOLUME_ NUMBER	VARCHAR2	20	Journals	Journal Volume Name	VARCHAR	400
Edition		EDITION_ NUMBER	VARCHAR2	255	Books	Book Edition Name	VARCHAR	400

Web		e-commerce and Customer System			PIM			
Book	**Journal**	**Attibute**	**Data Type**	**Data Size**	**PIM Entity**	**PIM Attribute**	**Data Type**	**Data Size**
	Issue Number	ISSUES_PER_ YEAR	DOUBLE		Journals	Issues per Year Quantity	INTEGER	38
		ELS_PRODUCT_ EN						
SubTitle	SubTitle	SUB_TITLE	VARCHAR2	255	Product	Product Sub Title	VARCHAR	400
Price-Format	Price-Format	WEB_ PRODUCT_ TYPE_NAME	VARCHAR2	100	Product	Purchase Type Code, Purchase Type Name	VARCHAR	35, 400
		DCSX_SKU						
Price-Format	Price-Format	SKU_TYPE	VARCHAR2	20	Product	Purchase Type Code, Purchase Type Name	VARCHAR	35, 400
		ELS_SKU						
Price-Format	Price-Format	PURCHASE_ TYPE	VARCHAR2	75	Purchase Type for Product, Product Purchase Type	Purchase Type Code, Purchase Type Name	VARCHAR	35, 400
		DCS_PRICE						
Price	Price	PRICE_LIST, DISPLAY_ NAME	VARCHAR2	40, 254	Product Price, Currency	Currency Code, Currency Name	VARCHAR	35, 400
Price	Price	LIST_PRICE	NUMBER (19,7)	12	Product Price	Price Amount	NUMERIC (15,2)	13
		ELS_PRODUCT_ REGIONAL_ INFO						
Publication Date		PUB_DATE	DATE	DATE	Product Lifecycle	Product Event Date	DATE	
Stock	Stock	PUB_STATUS	INTEGER	38	Product Lifecycle, Product Status	Status Code, Status Name	VARCHAR	35, 400
Short Description	Short Description	SHORT_ DESCRIPTION	CLOB	~	Product	Product Description	VARCHAR	3000
Long Description	Aim and Scope	LONG_ DESCRIPTION	CLOB	~	Product	Product Description	VARCHAR	3000
[drop down site]	[drop down site]	SITE_ID	VARCHAR2	40	Product Lifecycle, Region	Region Code, Region Name	VARCHAR	35, 400
		DCS_PRD_ PRNT_CATS						
Breadcrumb		CATEGORY_ID	VARCHAR2	40	Product Subject Area, Product	Product Subject Area Priority Code, Product Subject	VARCHAR	35, 400
Breadcrumb		DISPLAY_ NAME	VARCHAR2	254				

Web		e-commerce and Customer System			PIM			
Book	**Journal**	**Attibute**	**Data Type**	**Data Size**	**PIM Entity**	**PIM Attribute**	**Data Type**	**Data Size**
					Subject Area Priority	Area Priority Name		

Table 17 Unmapped Attributes for Integrated Applications and PIM

Web		e-commerce and Customer System		
Book	**Journal**	**Column Name**	**Data Type**	**Data Size**
e-commerce System				
		ANG_PRODUCT		
	Impact Factor	IMPACT_ FACTOR	NUMBER (10,3)	7
		DCSX_PRODUCT		
	5yr Impact Factor	VERSION_NUMBER	VARCHAR2	40
Table of Contents		TABLE_ OF_CONTENTS	CLOB	~
		ELS_PRODUCT_EN		
	Abstract and Indexing	ABSTRACT	CLOB	~
		ELS_DCSX_SKU		
		FULFILLMENT_ COMPANY_CODE	VARCHAR2	40
		ELS_SKU		
Stock	Stock	OUT_OF_PRINT	NUMBER (1)	1
		ELS_PRODUCT_ REGIONAL_INFO		
Key Feature		KEY_FEATURE	VARCHAR2	4000
Readership	Audience	AUDIENCE	CLOB	~
Review		QUOTES	CLOB	~
		FULFILLMENT_ COMPANY_CODE	VARCHAR2	40
Customer System				
CopyrightText	CopyrightText	CopyrightText	TEXT	255
SmallCover	SmallCover	SmallCover	TEXT	255
HomePageURL	HomePageURL	HomePageURL	TEXT	255
	VolumeIssueAlert Available	VolumeIssueAlert Available	TEXT	255
	Submityour ArticleURL	Submityour ArticleURL	TEXT	255
	SocietyURL	SocietyURL	TEXT	255
	SocietyText	SocietyText	TEXT	255

3.1.3 Conforming, Merging and Structuring the Schema

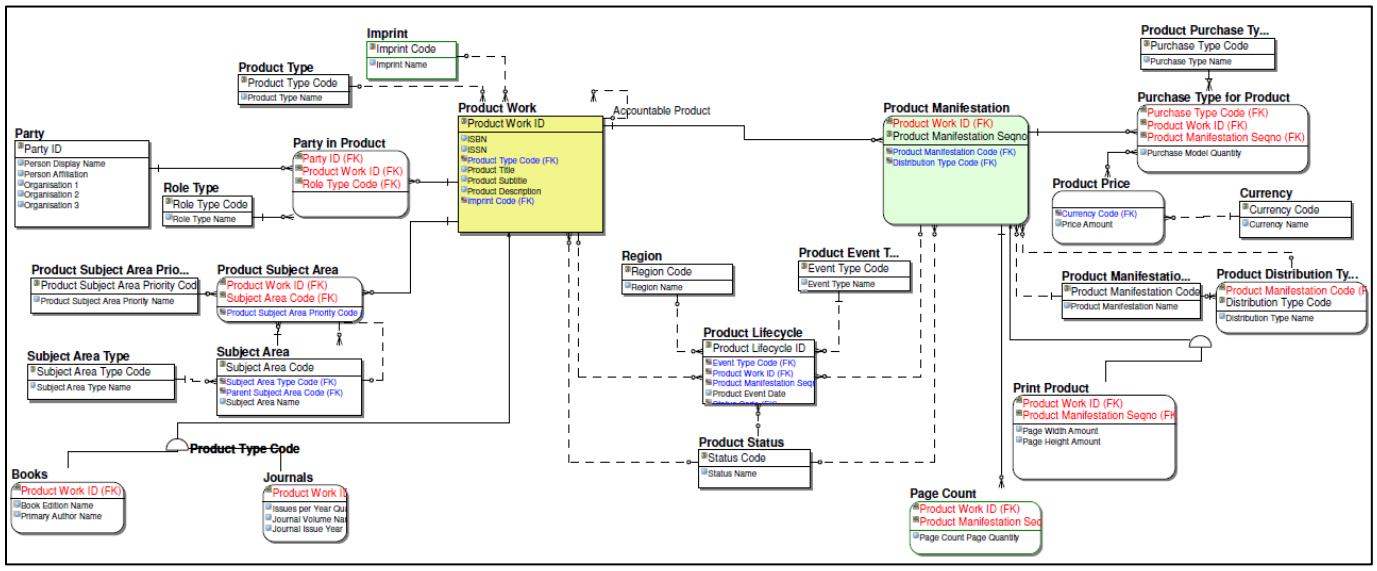

Figure 4 PIM Entities and Attributes

The end schema result should conform the existing PIM data model in Figure 4 because it represents the data model for several other applications. The table mapping between PIM and Integrated applications in Phase I is in Table 18.

Table 18 Entities - Tables Mapping

Entities	e-commerce Tables	Customer System Table	PIM Tables
Product Info	DCSX_PRODUCT, ANG_PRODUCT, ELS_PRODUCT, ELS_PRODUCT_EN, DCS_PRD_PRNT_CAT, and DCS_CATEGORY	CSRT (JOURNAL) and NONSERIALSRT (BOOK)	Product, Product Type, Books, Page Count, Print Product, Product Subject Area, Product Subject Area Priority, Journals, Party in Product, Party, Imprint
Product Regional Info	ELS_PRODUCT_REGIONAL_INFO	NA	Region, Product, Product Lifecycle, Product Status
Product SKU	DCSX_SKU, ELS_SKU, and ELS_DCSX_SKU	NA	Product, Purchase Type for Product, Product Purchase Type
Pricing List	DCS_PRICE, and DCS_PRICE_LIST	NA	Product Price, Currency

Within the mapped attributes in Table 2, there are several discrepancies that need to be resolved using the qualitative criteria as follows:

1. 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 criteria. 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 by other n (>2) applications in the company.

2. 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 value compound process could be conducted at the application level or data exchange.
3. 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. Elsevier could also revisit the product data model and opt to use international standard for conflict resolution, for example, using ONIX 3.0 specification[1].

Table 19 Conflict Resolution for Phase II Schema Integration

Web		Conflict Source		Resolution			Descripton
Book	**Journal**	**Column Name**	**Data Type/Size**	**Column Name**	**Data Type**	**Data Size**	
Title	Title	v	v	Product Title	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : PIM
Product Type	Product Type	v	v	Product Type Name	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : PIM
Print Book/ eBook ISBN	ISSN	v	v	ISBN/ ISSN	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : PIM
Author Name		**v**	v	Primary Author Name	VARCHAR	4000	▪ Column Name : PIM ▪ Data Type and Size : ES-CS
Number of Page		v	v	Page Count Page Quantity	INTEGER	38	▪ Column Name : PIM ▪ Data Type and Size : PIM
Dimensions		v	v	Page Height Amount, Page Width Amount	NUMERIC (15,2)	13	▪ Column Name : PIM ▪ Data Type and Size : PIM
Imprint		v	v	Imprint Name	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : PIM
Authors-Name	Editor Information	v	v	Person Display Name	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : ES-CS
	Editorial Information	v	v	Person Display Name	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : PIM
	Valume Number	v	v	Journal Volume Name	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : PIM
Edition		v	v	Book Edition Name	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : PIM

[1] http://www.editeur.org/8/ONIX/

Web		Conflict Source		Resolution			Descripton
Book	**Journal**	**Column Name**	**Data Type/Size**	**Column Name**	**Data Type**	**Data Size**	
	Issue Number	v	v	Issues per Year Quantity	INTEGER	38	▪ Column Name : PIM ▪ Data Type and Size : PIM
SubTitle	SubTitle	v	v	Product Sub Title	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : PIM
Price-Format	Price-Format	v	v	Purchase Type Name	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : PIM
Price	Price	v	v	Currency Name	VARCHAR	400	▪ Data Type and Size to PIM
Price	Price	v	v	Price Amount	NUMBER (19,7)	12	▪ Column Name : PIM ▪ Data Type and Size : ES-CS
Publication Date		v	v	Product Event Date	DATE	DATE	▪ Column Name : PIM
Short Description	Short Description	v	v	Product Description	CLOB	~	▪ Column Name : PIM ▪ Data Type and Size : ES-CS
Long Description	Aim and Scope	v	v	Product Description	CLOB	~	▪ Column Name : PIM ▪ Data Type and Size : ES-CS
[drop down site]	[drop down site]	v	v	Region Name	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : ES-CS
Breadcrumb		v	v	Product Subject Area Priority Name	VARCHAR	400	▪ Column Name : PIM ▪ Data Type and Size : ES-CS

While to conform the schemas for unmapped attributes in Table 17, we need to use correctness criteria. *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 PIM data model:

a. It is referenced in multiple business areas and business processes (Loshin [2]).
b. It is referenced in transaction and analytical system records (Loshin, [2]).
c. It tends to be static in comparison to transaction systems and do not change as frequently (Loshin [2]).
d. Low volume volatility (Otto, 2010).

Furthermore, to determine the importance of an attribute and to decide for adding the attributes in PIM, this study also use these methods:

a. Information comparison between e-store site and other prominent e-commerce site for Journal/ Book, for example, Barnes&Nobles, Amazon, SAGE, Wiley, and Springer.
b. Use data analytics approaches, for example, decision tree, cluster algorithm, or support vector machine, to develop hypotheses about important attributes for buying decision (Appendix 5).

Table 20 Add Attributes to PIM

Web		e-commerce System and CS			Desc	Condition			Add to PIM
Book	Journal	Column Name	Data Type	Data Size		Found in Other Sites	Found in 2 DA	Low Volatility	
	v	IMPACT_ FACTOR	NUMBER (10,3)	7	NA	v		v	Yes
	v	VERSION_NUMBER	VARCHAR2	40	NA	x		v	No
v		TABLE_ OF_CONTENTS	CLOB	~	NA	v	v	v	Yes
	v	ABSTRACT	CLOB	~	NA	x		v	No
		FULFILLMENT_ COMPANY_CODE	VARCHAR2	40	Should Exist	x		v	Yes
v		KEY_FEATURE	VARCHAR2	4000	NA	x	v	v	Yes
v	v	AUDIENCE	CLOB	~	NA	x	v	v	Yes
v		QUOTES	CLOB	~	NA	v	v	v	Yes
v	v	CopyrightText	TEXT	255	Generated using a string function	x		v	No
	v	VolumeIssueAlert Available	TEXT	255	No	x		v	No
	v	Submityour ArticleURL	TEXT	255	NA	x		v	No
	v	SocietyURL	TEXT	255	NA	x		v	No
	v	SocietyText	TEXT	255	NA	v		v	Yes

3.2 Data Quality Metrics Integration

3.2.1 Activities

This activity integrates the metrics specification for the new integrated schema. Because each attribute is attached to certain data quality metrics, we will describe the integration process from the metrics perspective. This process also uses the activities for data structure integration as follows:

3.2.1.1 Pre Integration

The pre integration activity is using the information from the data quality metrics development phase I and the product MDM data model (PIM) version 1.6.

3.2.1.2 Schema Comparison

This activity compares the metrics from both applications to find unmapped and mapped metrics as in Table 6. The PIM does not have data quality metrics specification, thus it should use all the metrics in the integrated applications.

Table 21 Metrics Availability Phase II

Metrics	e-commerce and Customer System	PIM
Completeness per Row		
CPR-01	Yes	NO

Metrics	e-commerce and Customer System	PIM
CPR-02	Yes	NO
CPR-03	Yes	NO
Syntactical Correctness		
SC-01	Yes	NO
SC-02	Yes	NO
Absence of Contradiction		
AOC-01	Yes	NO
AOC-02	Yes	NO
Absence of Repetition		
AOR-01	Yes	NO
Accuracy Inc. Currency		
ACR-01	Yes	NO
ACR-02	Yes	NO

3.2.1.3 Conforming, Merging, and Structuring the Schema

Within the mapped metrics in Table 6, there are several discrepancies that need to be resolved. The resolution change several attributes of DQ metrics in Table 7, namely data and rules in measurement methods. The resolution for the data quality metrics also uses the qualitative criteria as follows:

1. Different List of Attributes
   Using the completeness, we need to combine the attributes from both applications (See Conformed schema for mapped and unmapped attributes in section 3.1.3). This thesis uses all mapped attributes and only use unmapped attributes that are considered to add in PIM.
2. 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 attributes 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.

3.1 Result

3.1.1 Completeness

3.1.1.1 CPR-01

- Method : 1 - (Number of row with empty non-null able field divided with number of all row)
- Value : 0-1
- Attributes :
  - Table 19 in Resolution
  - Table 20 with add to PIM = Yes

3.1.1.2 CPR-02

- Method : 1 - (Number of unreferenced row (parentless row) divided with number of all rows)
- Value : 0-1
- Entities :
    - Parent : Product (ProductWork, Party, Product LifeCyle)
    - Child : ProductManifestasion (Product MAnifestasion, PurchaseTypeForProduct)

3.1.1.3 CPR-03

- Method : 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)
- Value : 0-1
- Entities :
    - Parent : Product (ProductWork, Party, Product LifeCyle)
    - Child : ProductManifestasion (Product MAnifestasion, PurchaseTypeForProduct)

3.1.1.4 CPR-04

result = 70%xCPR-01 + 15%xCPR-02 + 15%xCPR-03s

3.1.2 Syntactical Correctness

3.1.2.1 SC-01

- Method : 1 - (Number of row with non-standard value or format divided with number of all rows)
- Value : 0-1
- Basic Rules for Data Type
    - Text : "UNKNOWN", "EMPTY", "BLANK", "NULL"
    - Numeric : <=0
    - Date : 1 Jan 1900 or 1 Jan 1970
- Attributes and Rules :

a. Journal

Table 22 SC-01 for Journal (PIM)

Attributes	PIM	Rules
Info		
TITLE	Product[Product Title]	a. Basic Rule for Text b. All accented characters must be coded as Unicode Hexadecimal codes. E.g. the á should be encoded as á in the sheet; the & as & c. Please captitalize the first character of each noun and adjective, and of the very first word on the title. E.g. "the title of the new journal" becomes "The Title of the New Journal"
ISSN	Product[ISSN]	c. Text in e-commerce System d. remove dash (!). E.g. 12345678 and not 1234-5678.

Attributes	PIM	Rules
AUTHOR_ALIST	Party[Person Display Name]	Basic Rule for Text
AUTHOR_BLIST	Party[Person Display Name]	Basic Rule for Text
VOLUME_ NUMBER	Journals[Journal Volume Name]	Basic Rule for Text
EDITOR	Party[Person Display Name]	Basic Rule for Text
ISSUES_PER_YEAR	Journals[Issues per Year Quantity]	Basic Rule for Numeric
SUB_TITLE	Product[Product Sub Title]	Basic Rule for Text
IMPACT_FACTOR		Basic Rule for Numeric
Society Text		Basic Rule for Text
Regional Info		
SHORT_ DESCRIPTION	Product[Product Description]	Basic Rule for Text
AUDIENCE		Basic Rule for Text
Price		
LIST_PRICE	Product Price[Price Amount]	Basic Rule for Numeric

b. Book

Table 23 SC-01 for Books (PIM)

Attributes	**PIM**	**Type**
Info		
TITLE	Product[Product Title]	• Basic Rule for Text • The Title of the book without any reference to edition number/text or subtitle. • Glyphs that are non-standard (e.g. & and ©) or accented letters (e.g. Ü or ê) MUST be expressed in Unicode with this format : **&#x**<Unicode_value> • For example: & should be expressed as & (or &) © as © (or ©) Ü as ü (or ü) ê as ê (or ê)
ISBN	Product[ISBN]	Basic Rule for Text
AUTHOR	Product[Primary Author Name]	Basic Rule for Text
NUMBER_OF_	PageCount[Page	Basic Rule for Numeric

Attributes	PIM	Type
PAGES	Count Page Quantity]	
PUB_NUM_ LOG	Print Product[Page Height Amount, Page Width Amount]	Basic Rule for Numeric (for Print Journal/ Book)
TABLE_OF_ CONTENTS		Basic Rule for Text
AUTHOR_ALIST	Party[Person Display Name]	Basic Rule for Text
SUB_TITLE	Product[Product Sub Title]	• Basic Rule for Text • Glyphs that are non-standard (e.g. & and ©) or accented letters (e.g. Ü or ê) MUST be expressed in Unicode with this format : &#x<Unicode_value> • See **Error! Reference source not found.**: Unicode Charts for the most common ones.
Regional Info		
PUB_DATE	Product Lifecycle [Product Event Date]	Basic Rule for Date
SHORT_ DESCRIPTION	Product[Product Description]	Basic Rule for Text
KEY_FEATURE		Basic Rule for Text
LONG_ DESCRIPTION	Product[Product Description]	Basic Rule for Text
AUDIENCE		Basic Rule for Text
QUOTES		Basic Rule for Text
Price		
LIST_PRICE	Product Price[Price Amount]	Basic Rule for Numeric

3.1.2.2 SC-02

- Method : 1 - (Number of row with deviated value divided with number of all rows)
- Value : 0-1
- Attributes :

Table 24 SC-02 for Book and Journal (PIM)

Field	Book	Journal	PIM
Info			
PRODUCT_TYPE	v	v	Product [Product Type Code] Product Type[Product Type Name]
IMPRINT	v	v	Product[Imprint Code] Imprint[Imprint Name]
WEB_PRODUCT_ TYPE_NAME	v	v	PurchaseTypeForProduct[PurchaseTypeCode]

Field	Book	Journal	PIM
			ProductPruchaseType[Purchase Type Name]
CATEGORY_ID	v	v	Product Subject Area[SubjectAreaCode] SubjectArea[Subject Area Name]
Regional Info			
PUB_STATUS	v	v	ProductLifeCycle[Status Code] Product Status[Status Name]
FULFILLMENT _COMPANY _CODE	v	v	
SITE_ID	v	v	Product Lifecycle[RegionCode] Region [Region Name]
SKU			
SKU_TYPE	v	v	PurchaseTypeForProduct[PurchaseTypeCode] ProductPruchaseType[Purchase Type Name
FULFILLMENT _COMPANY _CODE	v	v	
PURCHASE_TYPE	v	v	PurchaseTypeForProduct[PurchaseTypeCode] ProductPruchaseType[Purchase Type Name
Price			
PRICE_LIST	v	v	Product Price[CurrencyCode] Currency[Currency Name]

3.1.2.3 SC-03

Result = average (SC-01, SC-02)

3.1.3 Absence of Contradiction

3.1.3.1 Rules

[1] Title - Category : Products with the same title have the same category

[2] Location - Price : The products region and price follow this matrix :

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

[3] Location&Format - Fulfillment Company Code :

- 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
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

3.1.3.2 AOC-01

- Method : 1 - (number of non-reasonable fields divided with number of all fields)
- Reasonable field : field that has the same top-5 values based on its distribution compared with previous data
- Value : 0-1
- Attributes :

Table 25 AOC-01 for PIM

Field	Book	Journal	PIM
Info			
PRODUCT_TYPE	v	v	Product [Product Type Code]
CATEGORY_ID	v	v	Product Subject Area[SubjectAreaCode]
Regional Info			
FULFILLMENT _COMPANY _CODE	v	v	
SITE_ID	v	v	Product Lifecycle[RegionCode]
SKU			
FULFILLMENT _COMPANY _CODE	v	v	
Price			
PRICE_LIST	v	v	Product Price[CurrencyCode]

3.1.3.3 AOC-02

- Method : 1 - (number of non-adhered business rules divided with number of all records)

- Business Rules : see 3.1.3.1
- Value : 0-1
- Attributes :

Table 26 AOC-02 for PIM

Field	Book	Journal	Rule	PIM
Info				
TITLE	v	v	1	Product [Product Title]
PRODUCT_TYPE	v	v	3	Product [Product Type Code]
CATEGORY_ID	v	v	1	Product Subject Area[SubjectAreaCode]
Regional Info				
FULFILLMENT _COMPANY _CODE	v	v	3	
SITE_ID	v	v	2,3	Product Lifecycle[RegionCode]
Sku				
FULFILLMENT _COMPANY _CODE	v	v	3	
Price				
PRICE_LIST	v	v	2	Product Price[CurrencyCode]

3.1.3.4 AOC-03

Result = average (AOC-01, AOC-02)

3.1.4 Absence of Repetition

3.1.4.1 AOR-01

- Method : 1 - (Number of duplicate row divided with number of all unique rows)
- Value : 0-1
- Attributes :
    a. Journal : Product[ISSN]
    b. Book : Product[ISBN]

3.1.5 Accuracy incl. Currency

3.1.5.1 ACR-01

- Method : average(number of unique book ISBN in PPM + number of unique book ISBN in French site XML/ number of unique book ISBN in PIM, number of unique ISSN in PROMIS/ number of unique ISSN in PIM)
- Value : 0-1
- Description : Not applicable for PIM because it's the source

3.1.5.2 ACR-02

- Method : 1- (min difference of time data is available with time data in PIM)/ 720
- Value : 0-1

- Description : Not applicable for PIM because it's the source

3.1.5.3 ACR-03

Result = average(CPR-01, SC-01, SC-02, ACR-01, ACR-02)

Appendix 1. Data Quality Metrics Specification for e-commerce System

Table 27 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 e-commerce 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 e-commerce system	Completeness per row	All in marketing data model
	▪ see (i) mapping problem				
iv	Internet user could	Potential revenue loss	Incomplete	Completeness per	All in websites

<table>
<tr><th>No</th><th>Business Problem</th><th>Business Impact</th><th>Data Defect</th><th>DQ Dimensions</th><th>Attribute</th></tr>
<tr><td></td><td>not find the data in top result using search engine</td><td></td><td>information in e-commerce system</td><td>row</td><td>data model</td></tr>
<tr><td></td><td colspan="5">▪ see (i) mapping problem</td></tr>
<tr><td rowspan="4">v</td><td rowspan="4">Offering unavailable product</td><td rowspan="4">Customer dissatisfaction, unrecognized revenue, ineffective marketing, and potential revenue loss</td><td>a. Inaccurate data in e-commerce system (Journal database)</td><td>Accuracy inc. currency</td><td>Saleable/ Availability in a Region</td></tr>
<tr><td>b. Incomplete data in e-commerce system (Journal database)</td><td>Completeness, Business Referential Integrity</td><td>Fulfillment system</td></tr>
<tr><td>c. Inconsistent data from Journal database and e-commerce system</td><td rowspan="2">Absence of contradiction, Accuracy incl. currency</td><td>Product data</td></tr>
<tr><td>d. Inaccurate data in e-commerce system</td><td>Product Data</td></tr>
<tr><td></td><td colspan="5">
▪ 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
<table>
<tr><th>Site</th><th>Currency</th></tr>
<tr><td>EST_UK_BS</td><td>GBP/ EUR</td></tr>
<tr><td>EST_AU_BS</td><td>USD</td></tr>
<tr><td>EST_ASIA_BS</td><td>USD</td></tr>
<tr><td>EST_US_BS</td><td>USD</td></tr>
<tr><td>EST_JP_BS</td><td>JPY/ USD</td></tr>
<tr><td>EST_MEA_BS</td><td>USD</td></tr>
<tr><td>EST_EU_BS</td><td>EUR</td></tr>
</table>
Book
<table>
<tr><th>Site</th><th>Currency</th></tr>
<tr><td>EST_UK_BS</td><td>GBP/ EUR</td></tr>
<tr><td>EST_AU_BS</td><td>AUD/ USD</td></tr>
<tr><td>EST_ASIA_BS</td><td>USD</td></tr>
<tr><td>EST_US_BS</td><td>USD</td></tr>
<tr><td>EST_JP_BS</td><td>JPY/ USD</td></tr>
<tr><td>EST_MEA_BS</td><td>USD</td></tr>
<tr><td>EST_EU_BS</td><td>EUR</td></tr>
</table>
4. TOTAL = 15%xCPR-01 + 15%xSC-02 + 70%xAOC-02
▪ Frequency : Daily
▪ Value : [0-1]
▪ Expected Threshold :
</td></tr>
<tr><td></td><td colspan="5">
▪ 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
</td></tr>
</table>

<table>
<tr><th>No</th><th>Business Problem</th><th>Business Impact</th><th>Data Defect</th><th>DQ Dimensions</th><th>Attribute</th></tr>
<tr><td></td><td colspan="5">
rule: Journal
<table>
<tr><th rowspan="2">Site</th><th colspan="2">Print Journal</th><th>eJournal</th></tr>
<tr><th>PJROMIS</th><th>PJARGI</th><th>EJSD</th></tr>
<tr><td>EST_AU_BS</td><td>DELTA</td><td>ARGI</td><td>CRM</td></tr>
<tr><td>EST_EU_BS</td><td>DELTA</td><td>-</td><td>CRM</td></tr>
<tr><td>EST_MEA_BS</td><td>DELTA</td><td>ARGI</td><td>CRM</td></tr>
<tr><td>EST_UK_BS</td><td>DELTA</td><td>-</td><td>CRM</td></tr>
<tr><td>EST_JP_BS</td><td>DELTA</td><td>-</td><td>CRM</td></tr>
<tr><td>EST_US_BS</td><td>DELTA</td><td>ARGI</td><td>CRM</td></tr>
<tr><td>EST_ASIA_BS</td><td>DELTA</td><td>ARGI</td><td>CRM</td></tr>
</table>
rule: Book
<table>
<tr><th rowspan="2">Site</th><th>Print Book</th><th colspan="2">eBook</th></tr>
<tr><th>Physical</th><th>EBSD</th><th>Others</th></tr>
<tr><td>EST_AU_BS</td><td>BOOKMASTER</td><td>CRM</td><td>DELTA</td></tr>
<tr><td>EST_EU_BS</td><td>DELTA</td><td>CRM</td><td>DELTA</td></tr>
<tr><td>EST_MEA_BS</td><td>DELTA</td><td>CRM</td><td>DELTA</td></tr>
<tr><td>EST_UK_BS</td><td>DELTA</td><td>CRM</td><td>DELTA</td></tr>
<tr><td>EST_JP_BS</td><td>COPS</td><td>CRM</td><td>DELTA</td></tr>
<tr><td>EST_US_BS</td><td>COPS</td><td>CRM</td><td>DELTA</td></tr>
<tr><td>EST_ASIA_BS</td><td>COPS</td><td>CRM</td><td>DELTA</td></tr>
</table>
4. TOTAL : 15%xCPR-01 + 15%xSC-02 + 70%xAOC-02
▪ Frequency : Daily
▪ Value : [0-1]
▪ Expected Threshold :
</td></tr>
<tr><td></td><td colspan="5">
▪ Data defect [c,d]
▪ Measurement Method :
1. ACR-01 : number of unique ISN in Journal database should be available for e-commerce/ number of unique ISN in e-commerce system -> min/ max, number of unique ISBN in Book database should be available for e-commerce / number of unique ISBN in e-commerce 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 :
</td></tr>
<tr><td>vi</td><td>Products are not included in the marketing campaign</td><td>Potential revenue loss</td><td>Taxonomy mapping problem</td><td>Absence of contradiction</td><td>Subject</td></tr>
<tr><td></td><td colspan="5">see (ii) mapping problem</td></tr>
</table>

Table 28 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) **result** = 1 - (Number of unreferenced row (parentless row) divided with number of all rows)	[0-1]	Daily	▪ String and 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	
	NOTE : ▪				
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	[0-1]	Daily	String and Numeric (deviation=0)

No	ID	Measurement Method	Value	Freq.	Attribute
		OR Jaro-Winkler distance>=0.8. ▪ Similarity=1 for numeric type field			
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; Consistent cross table multi columns profile:) ▪ Rules : Title - Category, Price - Location, Location/ Type - Fulfillment Company Code **result =** average (all ratio per avail)	[0-1]	Daily	String and Numeric
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
	NOTE: ▪				
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

No	ID	Measurement Method	Value	Freq.	Attribute
	NOTE: ▪				
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 e-commerce system, number of unique ISSN in Journal database/ number of unique ISSN in e-commerce 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 e-commerce system)/ 720 ▪ Ratio 2 : 1- (min difference of time data in e-commerce 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 2. Data Quality Metrics Specification for Customer System

29 Data Quality Metrics for Customer System

No	ID	Measurement Method	Value	Freq.	Attribute Type
Completeness per row (horizontal completeness)					
17	CPR-01	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)	[0-1]	Daily	▪ String and Numeric ▪ Required Attributes
	NOTE : ▪ CPR-01 also for Journal:SocietyURL-SocietyText case and Book:Editor-Author Case				
Syntactical correctness (conformity)					
18	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 based on distribution OR defined business rule (postcode is 4 char, dash, 2 numeric : ZZZZ-99)	[0-1]	Daily	▪ Numeric: between min-max value ▪ String and Numeric: business rule, string patter, top-3 value ▪ Non LoV attributes
19	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	▪ LoV attributes ▪ String and Numeric (deviation=0)
20	SC-03	**result** = Average (SC-01, SC-02) if there is no reference table for SC-02 then SC-03 = SC-01	[0-1]	Daily	String and Numeric (deviation=0)
	NOTE: ▪ SC-01 : 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 repetitions (free of duplicates)					
21	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
	NOTE: ▪ Is ISBN and ISSN a record identifier?				
	NOTE: ▪ ACR-01 : PPM and IGT are considered as the "Real World". In MDM, the repository holds the golden record..				

Rules for Syntactical Correctness

1) SC-01
   a) Journal

Table 30 SC-01 Rules for Journal

No	Field	Req.	Description (Rules for Correctness)
1	Name	Y	All accented characters must be coded as Unicode Hexadecimal codes. E.g. the **á** should be encoded as **á** in the sheet; the **&** as **&** Please captitalize the first character of each noun and adjective, and of the very first word on the title. E.g. “the title of the new journal” becomes “The Title of the New Journal”
2	Issues per year	Y	>0
3	Small Cover	Y	Must always be “S<issn>.gif, where <issn> is the unformatted ISSN – i.e. without hyphen – of the publication in question
4	ISSN	Y	Copy ISSN from IGT form, but remove dash (!). E.g. 12345678 and not 1234-5678.

   b) Book

Table 31 SC-01 Rules for Book

No	Field	Req.	Description (Rules for Correctness)
1	Title	Y	The Title of the book without any reference to edition number/text or subtitle. Glyphs that are non-standard (e.g. & and ©) or accented letters (e.g. Ü or ê) MUST be expressed in Unicode with this format : **&#x**<Unicode_value> For example: & should be expressed as & (or &) © as © (or ©) Ü as ü (or ü) ê as ê (or ê) See **Error! Reference source not found.**: Unicode Charts for the most common ones.
2	Subtitle	N	Glyphs that are non-standard (e.g. & and ©) or accented letters (e.g. Ü or ê) MUST be expressed in Unicode with this format : **&#x**<Unicode_value> See **Error! Reference source not found.**: Unicode Charts for the most common ones.
3	Formatted ISBN	Y	17 character with 4 hypen and 13 number : XXX-X-XX-XXXXXX-X
4	Small Cover	N	The SC value is made up of the unformatted ISBN with the suffix “_cov150h” and the file extension .gif. For example, 9780080444109_cov150h.gif

2) SC-02 (List of Values)
   a) Journal

No	Field	Value
1	Imprint Publisher	LoV Imprint
2	Sample Issue Available	Y/N
3	Journal/Book Alert Available	Y/N
4	Available Flag	Y/N

b) Book

No	OutputField	Value
1	Imprint Publisher	LoV Imprint
2	Available Flag	Y/N

Appendix 3. Discrepancy Assessment Results for SC-01

a. Journal

i. Attributes : Title

ii. Results

Table 32 e-commerce-Customer System Journal Data Discrepancies for SC-01

Title		
Records	Matched Records = 2,022	Total Records in e-commerce System = 2129
Discreprancy Number	249 (12.31%)	
Discrepancy Cause	Unicode (98 records)	Rule in CS
	Case Sensitive (15 records)	Rule in CS
	More texts in e-commerce System (13 records)	
	More texts in CS (16 records)	
	And / & (44 records)	html symbol
	SubTitle in Title	Rule in CS

b. Book

i. Attributes : Title, SubTitle

ii. Results

Table 33 e-commerce-Customer System Book Data Discrepancies for SC-01

Title		
Records	Matched Records = 8,686	Total Records in e-commerce System = 43,464
Discreprancy Number	1,106 (12.73%)	
Discrepancy Cause	Unicode (275 records)	Rule in CS
	Case Sensitive (163 records)	Rule in CS
	More texts in e-commerce System (171 records)	
	More texts in CS (25 records)	
	Edition/ Volume in Title (170 records)	Rule in CS
	SubTitle in Title	Rule in CS
Sub Title		
Records	8,686	Total Records in e-commerce System = 43,464
Discreprancy Number	2,768 (31.87%)	
Discrepancy Cause	Unicode (151 records)	Rule in CS
	Case Sensitive (642 records)	Rule in CS
	More texts in e-commerce System (1,223 records)	

	More texts in CS (584 records)	Could be a result of Subtitle in Title for e-commerce System data

Appendix 4. Discrepancy Assessment Results for SC-02

a. Journal

i. Attributes : Product Type, Imprint

ii. Results

- Product_Type
  - List of Values Discrepancies

e-commerce System	Customer System
Journal	▪ Handbook Series ▪ Journal ▪ Commercial Book Series

  - Data Discrepancies : <None>
- Imprint
  - Data Discrepancies :
    - Records : 2,022 (Total Records in e-commerce System = 2129)
    - Discepancy : 930 (46.02%)
    - Top-2 incorrect mappings:
      1. e-commerce System (ELSEVIER) -> CS (ANY) : 716 records
      2. e-commerce System (ANY) -> CS (Elsevier) : 85 records

Table 34 Journal Imprint Discrepancies

Imprint Discrepancy (e-commerce -> Customer)	Count
Australian Physiotherapy Association->Elsevier	1
BAILLIÈRE TINDALL->Baillière Tindall	1
Chinese Society for Metals->Elsevier	1
CURRENT OPINION->Elsevier	2
CURRENT OPINION->Elsevier Current Trends	9
Ei - ENGINEERING INFORMATION->Elsevier	4
ELSEVIER ADVANCED TECHNOLOGY->Elsevier	2
ELSEVIER MASSON->Elsevier	2
ELSEVIER URBAN & PARTNER->Elsevier	4
ELSEVIER URBAN & PARTNER->Urban & Fischer	4
ELSEVIER->Academic Press	156
ELSEVIER->Baillière Tindall	5
ELSEVIER->Churchill Livingstone	17
ELSEVIER->Content Repository Only!	1
ELSEVIER->Elsevier Advanced Technology	7
ELSEVIER->Elsevier Current Trends	1
ELSEVIER->Elsevier Doyma	5
ELSEVIER->Elsevier Masson	71
ELSEVIER->Elsevier Science B.V.	1
ELSEVIER->Excerpta Medica	1

Imprint Discrepancy (e-commerce -> Customer)	Count
ELSEVIER->JAI	27
ELSEVIER->Mosby	5
ELSEVIER->No longer published by Elsevier	5
ELSEVIER->North-Holland	82
ELSEVIER->Pergamon	266
ELSEVIER->Urban & Fischer	49
ELSEVIER->W.B. Saunders	17
EXCERPTA MEDICA->Elsevier	1
HANLEY AND BELFUS MEDICAL PUBLISHERS->Elsevier	1
HANYANG UNIVERSITY->Elsevier	1
MASSON->Elsevier Masson	29
MEDICINE PUBLISHING->Churchill Livingstone	2
MEDICINE PUBLISHING->Elsevier	4
MOSBY->W.B. Saunders	1
NORTH-HOLLAND->Elsevier	2
Northwest Institute for Nonferrous Metal Research->Elsevier	1
SAUNDERS->Elsevier	55
SAUNDERS->W.B. Saunders	68
SCIENCE PRESS->Elsevier	2
TRENDS->Elsevier	1
TRENDS->Elsevier Current Trends	15
URBAN AND FISCHER->Elsevier	1
Grand Total	**930**

b. Book

i. Attributes : Product Type, Imprint

ii. Results

- Product_Type
  - List of Value Discrepancies

e-commerce System	Customer System
▪ Ebook	▪ eBook
▪ Printbook	▪ Major Reference Work

  - Data Discrepancies
    - Records : 8,686 (Total Records in e-commerce System = 43,464)
    - Discepancy : 8,507 (97.94%)
    - Most Incorrect mapping :
      1. e-commerce System (Printbook) -> CS (eBook) : 8,462 records (97.42%)

Table 35 Book Product Type Discrepancies

e-commerce System	Customer System	Records
Ebook	Major Reference Work	26
Printbook	Major Reference Work	19
Printbook	eBook	8,462

- Imprint
  - Data Discrepancies
    - Records : 8,686 (Total Records in e-commerce System = 43,464)
    - Discepancy : 2,343 (26.97%)
    - Top-3 incorrect mappings:
      1. e-commerce System (ELSEVIER) -> CS (ANY) : 716 records
      2. e-commerce System (ANY) -> CS (Elsevier) : 364 records
      3. e-commerce System (ANY) -> CS (Content Repository Only!) : 966 records

Table 36 Book Imprint Discrepancies

Imprint Discrepancy (e-commerce -> Customer)	Count
ACACL->Academic Press	1
ACACL->Content Repository Only!	1
Academic Press->American Physiological Society	2
Academic Press->Content Repository Only!	8
Academic Press->Elsevier	17
Academic Press->Elsevier Science	1
Academic Press->William Andrew Publishing	1
Anderson->Anderson Publishing, Ltd.	56
Bailliere Tindall->Content Repository Only!	2
BH/Optician->Butterworth-Heinemann	1
Butterworth-Heinemann->Academic Press	1
Butterworth-Heinemann->Architectural Press	1
Butterworth-Heinemann->Arnold	1
Butterworth-Heinemann->Content Repository Only!	4
Butterworth-Heinemann->Elsevier Science	1
Butterworth-Heinemann->Kogan Page Science	7
Butterworth-Heinemann->Newnes	5
ChemTec Publishing->Elsevier	6
Churchill Livingstone Australia->Content Repository Only!	24
Churchill Livingstone->Content Repository Only!	100
Churchill Livingstone->Elsevier	2
Churchill Livingstone->Mosby	1
Digital Press->Butterworth-Heinemann	1
Elsevier Science->Academic Press	5

Imprint Discrepancy (e-commerce -> Customer)	Count
Elsevier Science->Butterworth-Heinemann	4
Elsevier Science->Content Repository Only!	8
Elsevier Science->Elsevier	330
Elsevier Science->Elsevier Science B.V.	79
Elsevier Science->Elsevier Science Ltd	96
Elsevier Science->North-Holland	1
Elsevier Science->Pergamon	6
Elsevier->Academic Press	3
Elsevier->Butterworth-Heinemann	1
Elsevier->Content Repository Only!	10
Elsevier->Elsevier Science Ltd	1
Gulf Professional Publishing->Butterworth-Heinemann	4
Hanley & Belfus->Content Repository Only!	2
Hanley & Belfus->Hanley & Belfus	4
Hanley & Belfus->Mosby	2
Hanley & Belfus->W.B. Saunders	1
JAI Press->Elsevier Science	1
JAI Press->JAI	2
JAI Press->North-Holland	1
Made Simple->Academic Press	1
Morgan Kaufmann->Academic Press	5
Mosby Australia->Content Repository Only!	10
Mosby Canada->Content Repository Only!	1
Mosby Ltd.->Content Repository Only!	11
Mosby Ltd.->Mosby	12
Mosby/JEMS->Content Repository Only!	5
Mosby->Content Repository Only!	191
Mosby->W.B. Saunders	2
Newnes->Butterworth-Heinemann	7
North Holland->Elsevier	6
North Holland->Elsevier Science B.V.	1
North Holland->North-Holland	38
Pergamon->Academic Press	1
Pergamon->Elsevier	3
Pergamon->Elsevier Science	2
Pergamon->Elsevier Science Ltd	2
Saunders Australia->W.B. Saunders Australia	3
Saunders Ltd.->Content Repository Only!	37
Saunders Ltd.->W.B. Saunders	50
Saunders->Churchill Livingstone	1
Saunders->Content Repository Only!	548

Imprint Discrepancy (e-commerce -> Customer)	Count
Saunders->Hanley & Belfus	1
Saunders->W.B. Saunders	317
URBFI->Churchill Livingstone	1
URBFI->Content Repository Only!	4
William Andrew->William Andrew Publishing	278
Grand Total	**2343**

Appendix 5. Data Analytics to Select Important Attributes

1. Overview

Data Analytics activity is conducted to select important attributes to e-commerce in Elsevier. Furthermore, the results could be used as a reference to add unmapped attributes to Product MDM data model (PIM) or for data quality improvement activities, e.g., which attributes should have complete information. Because this is only additional, this thesis only provides simple data analytics with these assumptions:

a. The quality dimension is only related to completeness
b. The data is only for attribute in the Product Info entity. The assessment is only to select incomplete attributes. Complete attributes (>=95% completeness) are considered important
c. The important attributes are the one used for buying decision. Thus, the data set is analyzed with visits and sales data in 2013.

2. Data Sets
   a. There are three data sets in this assessment, namely Book Sales Data in 2013, Bok Visits Data in 2013 from Google Analytics, and Book Completeness Assessment data from e-commerce data.
   b. The attributes for completeness are for Product Info as in Table 37 and Table 38.

Table 37. Book Info CPR-01 Assessment

Column Name	Criteria	OK	Completeness %
PRODUCT_ID	Not Blank and Not Null	43,464	100.00
TITLE	Not Blank and Not Null	43,464	100.00
PRODUCT_TYPE	Not Blank and Not Null	43,464	100.00
ISBN	Not Blank and Not Null	43,464	100.00
ALL_AUTHOR	Not Blank and Not Null	43,108	99.18
NUMBER_OF_PAGES	Not Blank and Not Null	40,016	92.07
PUB_NUM_LOG	Not Blank and Not Null	13,327	30.66
TABLE_OF_CONTENTS	Not Blank and Not Null	29,718	68.37
IMPRINT	Not Blank and Not Null	43,464	100.00
NEXT_EDITION_ISBN	Not Blank and Not Null	2,263	5.21
AUTHOR_ALIST	Not Blank and Not Null	43,413	99.88
SUB_TITLE	Not Blank and Not Null	14,061	32.35
WEB_PRODUCT_TYPE_NAME	Not Blank and Not Null	41,518	95.52
CATEGORY_ID	Not Blank and Not Null	43,464	100.00

Table 38 Book Regional Info CPR-01 Assessment

Column Name	Criteria	OK	Completeness %
SITE_ID	Not Blank and Not Null	292,647	100.00
PUB_DATE	Not Blank and Not Null	201,488	68.85
PUB_STATUS	Not Blank and Not Null	292,647	100.00

Column Name	Criteria	OK	Completeness %
SHORT_DESCRIPTION	Not Blank and Not Null	149,420	51.06
KEY_FEATURE	Not Blank and Not Null	158,548	54.18
LONG_DESCRIPTION	Not Blank and Not Null	169,122	57.79
AUDIENCE	Not Blank and Not Null	129,000	44.08
QUOTES	Not Blank and Not Null	65,021	22.22
FULFILLMENT_COMPANY_CODE	Not Blank and Not Null	292,647	100.00

c. Determine the importance of these attributes: PUB_NUM_LOG, TABLE_OF_CONTENTS, NEXT_EDITION_ISBN, SUB_TITLE, PUB_DATE,SHORT_DESCRIPTION, KEY_FEATURE, LONG_DESCRIPTION, AUDIENCE, QUOTES

3. Data Analytics Activity and Result
   a. Tool : Weka version 3.6[2]
   b. Total Records : 31,949 (81%=NO, 19%=YES)
   c. Feature
      - Cross Validation : 3 folds
      - CostSensitive Classiefier to reduce the False Negative (FN) because the data sets is imbalanced

Table 39 Confusion Matrix

		predicted	
		y	n
actual	y	TP	FN
	n	FP	TN

   d. Methods
      i. ADTree : select the attributes that lead to buying decision = YES
      ii. Decision Table : select the attributes that are used in the decision process
      iii. Genetic Algorithm : select the attributes that contribute forthe process
      iv. SVM : select the attributes with acceptable (mean + standard deviation) merit values
   e. Results

Table 40 Data Analytics Result

Method	Attributes	Description
ADTree	Sub Title, Short Description, Pub Date, Audience, Quotes	Correct Prediction Rate : 73-75%, TP 50-80%, FP 20-45%
Decision Table	Table Of Content, Dimension, Sub Title, Short Description, Pub Date, Key Feature , Audience, Quotes	
Genetic Alg.	Short Description, Pub Date, Key Feature, Audience,	NA

[2] http://www.cs.waikato.ac.nz/ml/weka/

	Quotes	
SVM	Table Of Content, Dimension, Key Feature, Audience, Quotes	NA

f. Attributes that are considered important (featured in 2 methods) : Sub Title, Short Description, Table Of Content, Key Feature, Audience, Quotes, Pub Date

4. Note
   - The correct prediction rate for the classifier (ADTree, Decision Table) is only 75% while it is expected to be more than 90% with low False Positive. Other data analytics methods also provide similar result. However this could be used as a guide for the next activities in Elsevier using data sets with a longer duration.

Appendix 6. Workshops Documents

References

[1] 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.

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

[3] 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.