Phase I: Identification
Business Problems and Data Defects

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 Figures

List of Tables

Document Control

Document Version : 1.0

Version history

Version	Final/Draft	Release date	Contributor	Comments
1.0	Final	21-May-2014	Marhendra L	▪ This document is developed using Document Assessment and Interviews Results in Elsevier ▪ This is a reference document for Main Report ▪ This document describes the phase I process and results

1 Information Gathering and Development Process

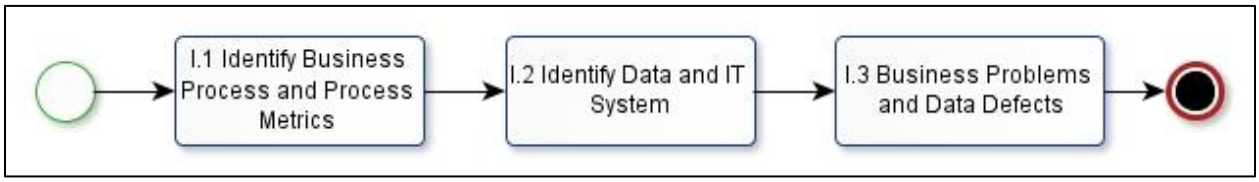

Figure 1 Phase I Identification, Otto, et al [1]

The main goal of this phase is to identify the business problems and the causing data defects. On the basis of the process model in Figure 1, there should be 3 different activities in phase I where those activities are conducted in serial. This is not the case for this study in Elsevier because of these conditions:

a. The documentation of business processes and IT systems is not complete
b. The knowledge of all the related processes and IT systems is dispersed into several people or teams
c. Time effectiveness and efficiency is crucial as the nature of a project

Those conditions introduce some adjustments to the process model, for example, by having parallel activities and some process breakdowns to iterate. In this thesis work, the activity I.1 and activity I.2 are conducted in parallel where each activity will have these several sub activities: Preliminary Interview Document to capture the components required in the meta-model, Interview Document Preparation to capture the essential information about the required component, Interview Process, Follow-ups, and Validation/ Confirmation. These set of sub activities are applicable for each information component in the Table 1.

Table 1 Process - Information Matrix for Phase I

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 defect						

Information	Corporate data steward	Client	Process owner	Data users	Senior technical data steward	Technical data steward
Business Goal. Impairment	x		x			
Business Problems. Cause	x			x	x	x
Defect data	x			x	x	x
Preventive measure (all)	x			x	x	x
Reactive measure (all)	x			x	x	x

Some of the information for the I.3 are also developed within the activity I.1 and I.2 and make the three activities seemed conducted in parallel. However, the information for I.3 is developed after the information for I.1 and I.2 are gathered/ developed.

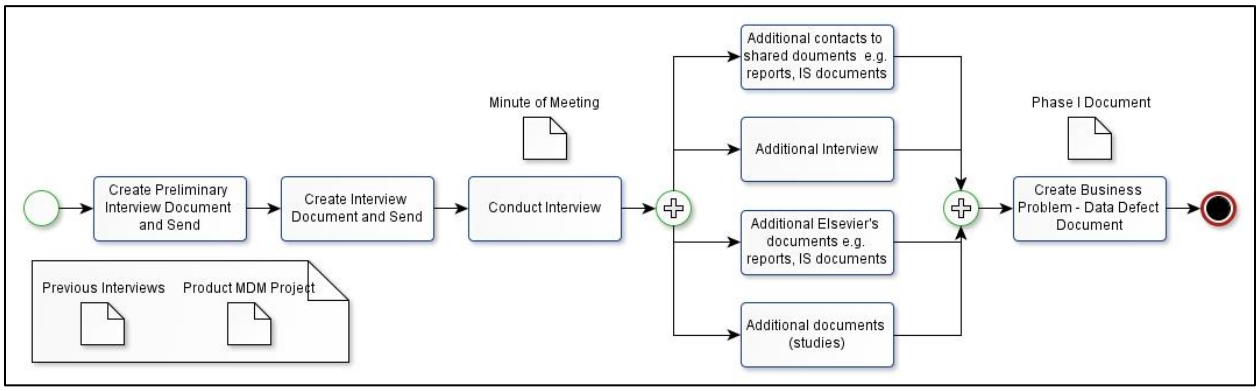

Figure 2 Phase I Process in Elsevier

The activities that are conducted during the phase I Identification (Figure 2) consist of literature study, document review, and interviews. The activities are as follows:

a. Preparing the Interview Document for Business Process, IT System, and Business Problem-Data Defect.

The activity I.1 and I.2 goals are to identify business process and process metrics for the remaining identification process and to identify IT systems (e.g. ERP systems, CRM systems, or databases) that are supporting the identified business processes. These activities also aim to list the potential interviewees for activity I.3. Because there are impediments as previously described, the interview documents are developed as follows:

i. There are 2 types of document for the interview process as follows:
   - The preliminary document to capture the existing landscape of the business and IT, and
   - The interview document to obtain the business problems and data defects, and to validate the important or unclear information in the preliminary document.

ii. The Business process

- Preliminary document

  Preliminary document is used to obtain information about several components in the meta-model (Table 1), namely business processes, accountability, performance metrics/ process indicator, business goal, and the existing efforts to maintain data quality (preventive and reactive measures). Some examples are provided to help the interviewee to understand the question and possible answers.

- Interview document

  The interview document is used to obtain information about some components in the meta-model (Table 1), namely business problems and their impacts (business goal impairment), the possible causing data defects.

  A set of data quality dimensions developed by Morbey [3] in Table 15 are used within the interview to provide common definition of data quality in this study.

iii. The IT systems

- Preliminary document

  Preliminary document is used to obtain and validate the information about current IT landscape that represents several components in the meta-model (Table 1), namely business applications, data (data exchange, data rules, and product data model), and the modules in applications to maintain the data quality (preventive and reactive measures).

- Interview document

  The interview document is used to get information about known data defect and existing efforts in IT to maintain data quality (preventive and reactive measures).

  A set of data quality dimensions mapped to the dimensions by Morbey [3] in Table 15 are used within the interview. Its aim is to provide the standard definition of data quality in this study. This is also useful to determine the data quality dimensions and data quality metrics in phase II.

b. Conducting the interview

The interviews in Appendix 1 were conducted in 2013 and the interviewees were Business and IT personnel. The goals of the interview are to validate the information in preliminary documents and to get the main result of phase I, namely known business problems and possible data defects.

c. Follow-ups

The follow-ups are important to provide flexibility to the process by allowing the interviewee to provide the details after the interview process and to introduce a new set of iterative activity (a-b) to get the complement information.

d. Validation/ Confirmation

The obtained or developed information is confirmed/ validated using the minutes of meeting. A meeting to confirm the developed business process and data defect is also conducted to validate the content. The result of document assessment and interview activities is the information in the next sections.

2 E-commerce in Elsevier

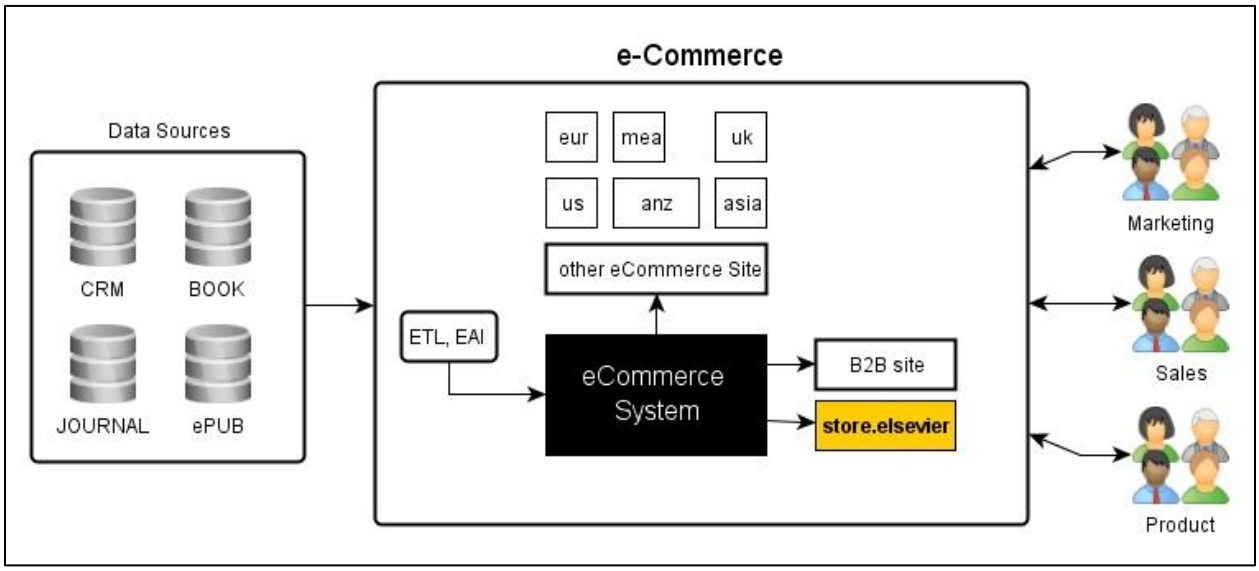

Figure 3 e-commerce System Context Diagram

The e-commerce system serves several e-commerce sites, but this study limits only to e-store (e-store.elsevier.com). The e-commerce system uses internal component to update the product catalogue using a batch process and manual process. In importing the data that is conducted daily, the batch process reads the XML files provided by the other systems. The e-commerce system also conducts a data quality check during the data import. The e-commerce system has several main data sources: Book database for books, Journal database for journal, CRM Database for electronic subscription, and ePub database for the electronic format product. The e-commerce system is internally operated by Product business unit to serve the Sales and Marketing business unit. The marketing is also using e-commerce system for campaigns.

3 Back Office Processes in e-commerce

In a complete e-commerce solution we could find several processes from product information availability, marketing, logistic, customer order and payment, and delivery. This study will only focus product information availability and marketing (Figure 4) because the two processes are highly related to product data.

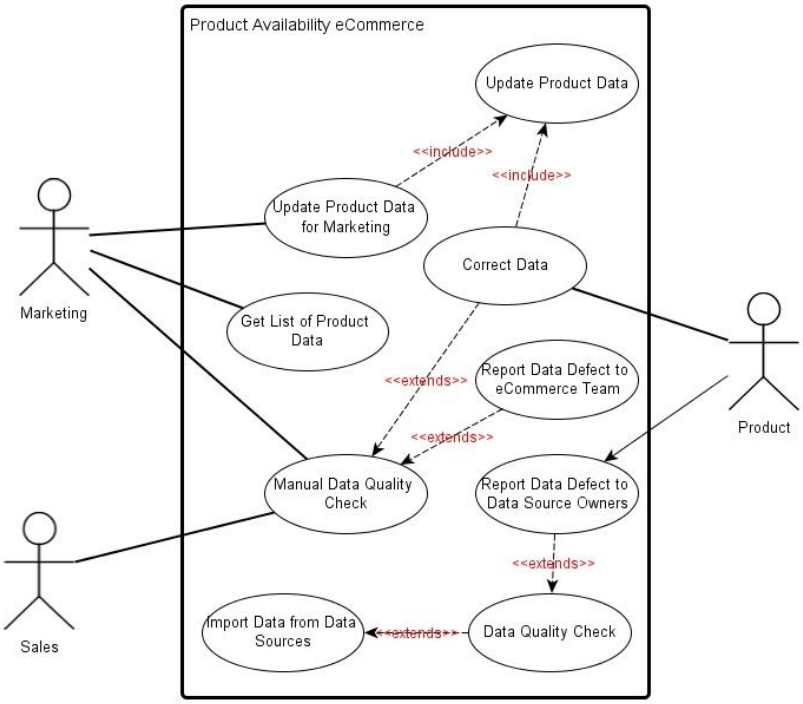

Figure 4 e-commerce UseCase

3.1 Product Information Availability

This process has a goal to provide product data for publishing in e-commerce site. The activities within this process (Figure 5) are data gathering from several source systems, data quality check, and data publishing. In this process, the product data is imported from several sources using the Extract, Transform, Load (ETL) and the Enterprise Application Integration (EAI) product. It is also mapped into an e-commerce product data model during the import process. The data quality check activities are conducted in 2 types of process, namely automatic process within the data import processes and manual process by sales/ marketing staff using a modul in the e-commerce application.

The activities within this process use all attributes of product data and could make an update of some attributes or an entity (e.g. delete). This process has uses cases as described in Table 2.

Table 2 Use Cases for Product Availability

No	Use Cases	Extends/ Includes	Actor	Attributes	Description
1	Import Data from Data Sources	Check Data Quality	System	All in Database	Data import from data sources is conducted automatically
2	Check Data Quality	Report Data Defect	System	All in	Check data quality is part of

No	Use Cases	Extends/ Includes	Actor	Attributes	Description
		to Data Source		Database	data import process
3	Report Data Defect to Data Source		Product and/ or e-commerce system team	All in websites	Error logs from import process that inform data quality defect will be reported back to the source system
4	Manual Data Check	Report Data Defect to e-commerce, Correct Data	Sales, Marketing	All in websites	This is an unscheduled and manual activity conducted by the e-commerce users (internal)
5	Report Data Defect to e-commerce		Sales, Marketing	All in websites	The sales/ marketing personnel could report to e-commerce team if a defect is found.
6	Correct Data	Update Data	Sales, Marketing, Product	All in websites	The personnel could fix/ update the data if a defect is found.

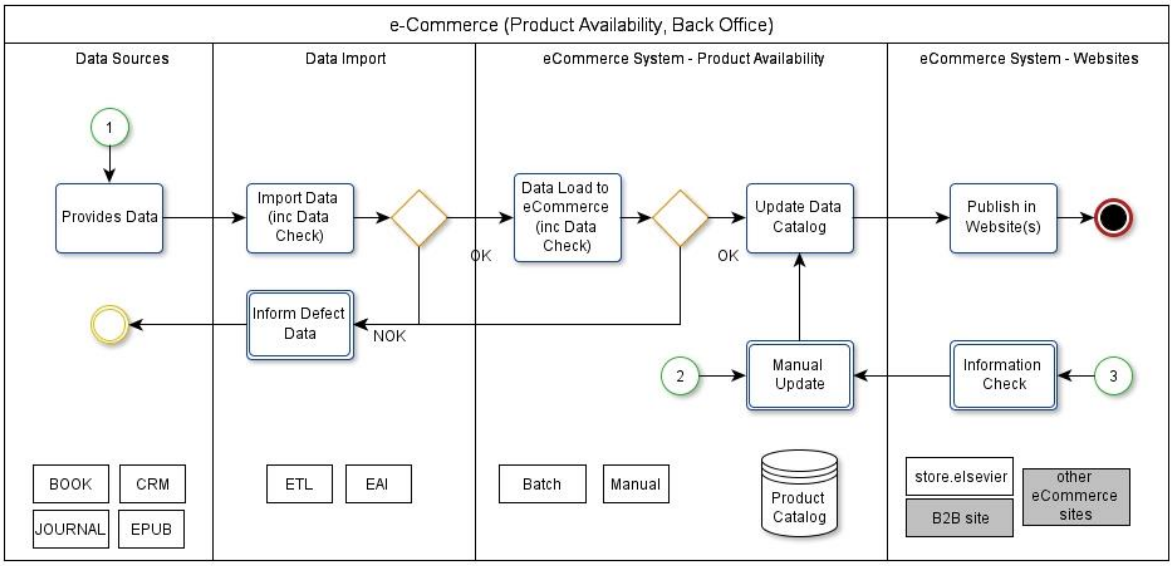

Figure 5 Product Availability Process

3.2 Marketing

This process has a goal to increase sales in e-commerce system by conducting a marketing campaign using electronic channels like Google AdWords and email. The marketing process has several activities:

- Planning

  This activity includes defining the marketing time period, product types (Global Campaign/ All product types or subject oriented campaign), discount rate (with discount or raising awareness), and marketing channel (email or Google AdWords)

- Campaign

  This activity includes obtaining the product information from the e-commerce system data feed, providing product information to 3rd parties (email, Google AdWords), and run the campaign as planned using Google AdWords, and landing pages in e-commerce system.

- Assessment

  Assessment of marketing performance is using Google AdWords dashboard (inc. Google Analytics).

  The planning and campaign activities use several attributes of product data for AdWords[1] and email.

This process has uses cases as described in Table 3 and Figure 6.

Table 3 Uses cases for Marketing

No	Use Cases	Extends/ Includes	Actor	Attributes	Description
1	Update product data for marketing process	Update product data	Marketing	Price, promotional code	Marketing staff adjusts the price and promo information for some products in a marketing campaign
2	Get list of product data		Marketing	All for campaign	Marketing staff gets a list of product data from e-commerce system to be published in marketing channels

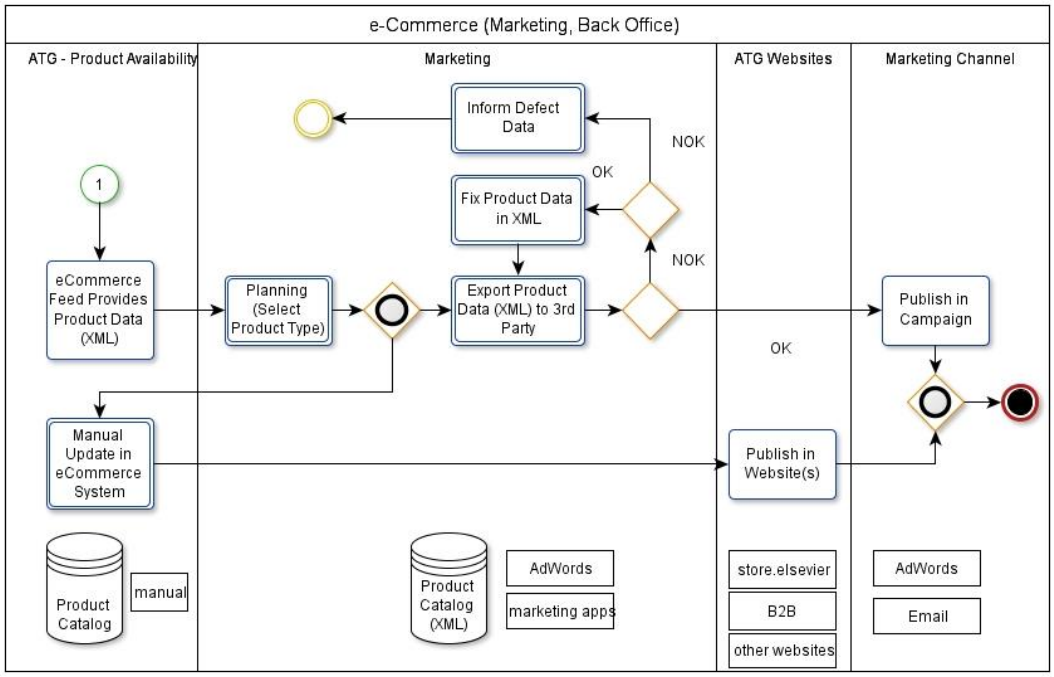

Figure 6 Marketing Process

3.3 Performance Metrics

i. Key Performance Indicator (KPI)

The goal of the product availability process is to update the product catalog in e-commerce system with the changes in the source systems with minimal latency (1 day). While the goal of marketing is to exceed the sales forecast that has been set at the beginning of each year. The e-commerce site (e-store) also has several performance metrics related to e-commerce as follows:

- Acquisition : Increase exposure/traffic to 1.9m visitors per year
- Conversion : Increase conversion rates to 1.1% in 2012

[1] https://support.google.com/merchants/answer/188494

- AOV : Achieve average order size of $96.04
- Retention : Increase repeat visitors through loyalty offers
- Stickiness : Improve UX (User Experience) to reduce bounce rate to <50%

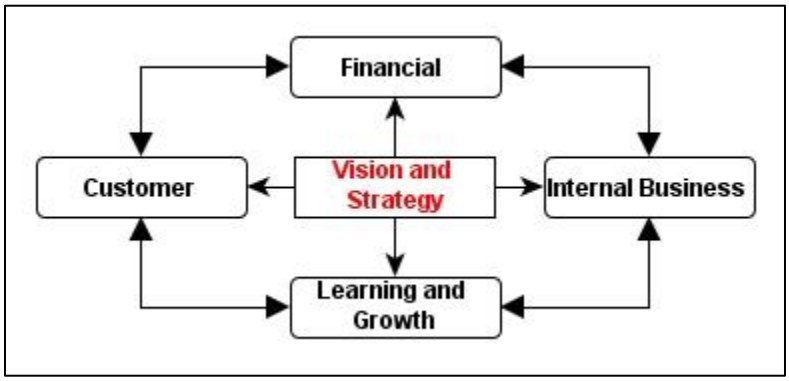

Figure 7 BSC Framework

Because the KPI for the activities in Elsevier e-commerce has not been formulated, we will develop the metrics on the basis of other studies in e-commerce. Tsai, et al. [5] developed e-commerce metrics (KPI) using a Balanced Scorecard Framework (BSC, Figure 7) and an empirical study by interviewing the experts. The result of the interview is processed using Delphi method and the selected metrics are as described in Table 11. We select the components in Balanced Score Card and metrics as follows:

a. The BSC by Kaplan describes that customer and internal business component have direct impact to financial component. This is the case for learning and growth. Pearlman [5] provided two models for path analytics in Figure 8 as follows:
   - Model 1
     The difference with BSC is that the learning and growth does not directly affect the customer. The components that directly affect financial are the customer and internal business.
   - Model 2
     In this model, learning and growth also directly affect the financial and it has a larger coefficient compared with customer and internal business. However, the subcomponent in learning and growth, which is a new product introduction, is not related to product data or record quality.

   The case by Tsai, et al. [5] using DEMATEL also provides that customer is the most impacted components and internal business is the component that gives most impact. On the basis of those studies, we only **select the metrics under customer and internal business component**.

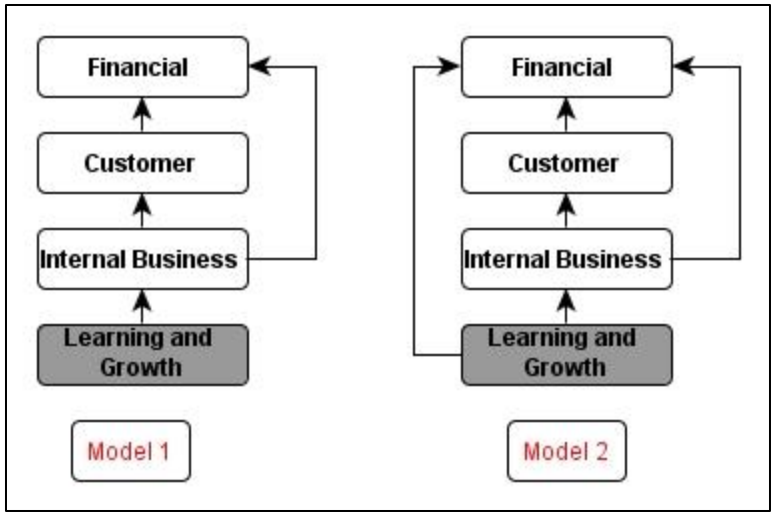

Figure 8 BSC Component Relationship, Perlman [5]

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. We select **only metrics that are related to product data** as in Table 4.

Table 4 Selected Metrics for e-commerce, Tsai, et al. [6]

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 Data Model for Elsevier e-commerce

There are 3 data models that are reviewed within this study: data model in the product catalogue repository (e-commerce system), data model in web sites, and data model for a marketing campaign. Within the e-commerce system, data models in websites and marketing campaign are the models that are used by the end user e.g. web users or customers. The data model in the product catalogue repository should be complete enough to meet the requirements of the other 2 data models.

i. Data model in web sites

The complete data model for the Elsevier e-commerce web site (e-store) can be found within e-store documentation. This data model is used in the presentation layer or web UI in e-commerce sites. The entity diagram for the Journal is Figure 9 and for the Book is Figure 10. A comparison of information about internal e-commerce (HS, myElsevier, and e-store) and competitors is described in Appendix 3. We could conclude that the information within the e-commerce site data model as follows:

- e-commerce sites in Elsevier (e-store) have more complete information compared with competitors (Amazon, Barnes and Nobles, SAGE, Wiley).
- There are several minor attributes that could be added to the Elsevier sites, for example file size and shipping weight.
- Elsevier store (e-store) website has more complete information compared with other internal e-commerce sites. The data model used by e-store is considered as the product data model required by the e-commerce system.

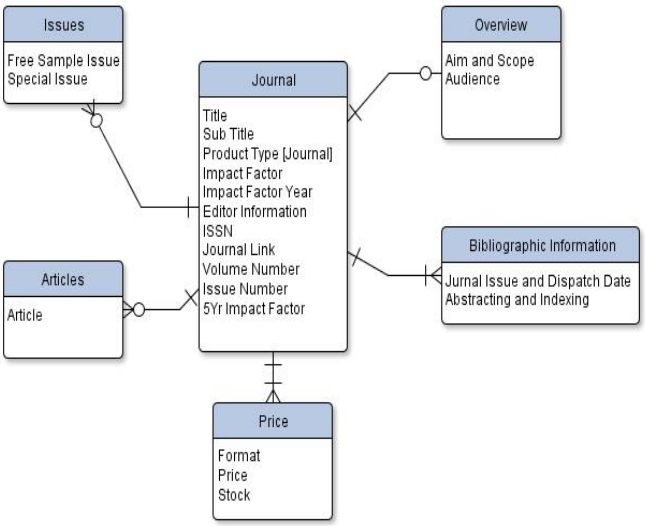

Figure 9 Journal Data Model in e-store Website

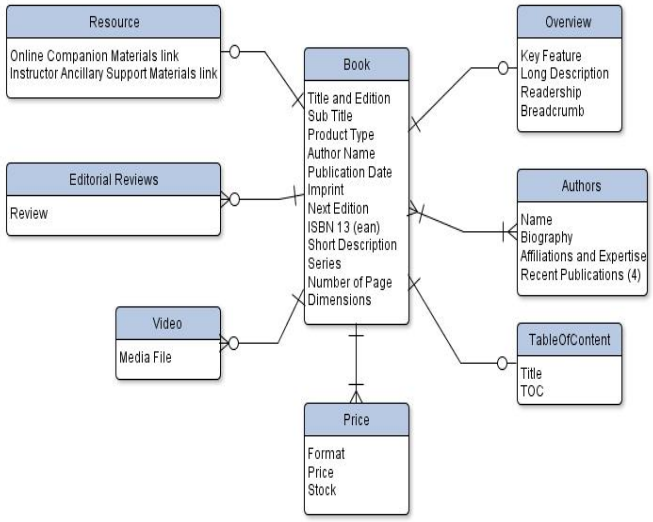

Figure 10 Book Data Model in e-store Website

ii. Data model for marketing campaign

There is also another data model that is used for marketing with AdWords channel as in Figure 11. This is also an important data model because its content will be used by Google to respond their users' book/ journal search activities.

Figure 11 Product data model for Marketing

iii. Data model in in product catalogue repository (e-commerce system) and its mapping with other data models

The complete data model for product in e-commerce is as in Table 14. There is only 1 schema for all product types (Books, Journal, Conferences, Others) and it is composed of 8 main entities: Product, CustomProduct, ElsProductRegionalInfo, Stock Keeping Unit (SKU), CustomSKU, PricingDetail, InventoryDetail, and ElsLinkedFormatMulti (Figure 12).

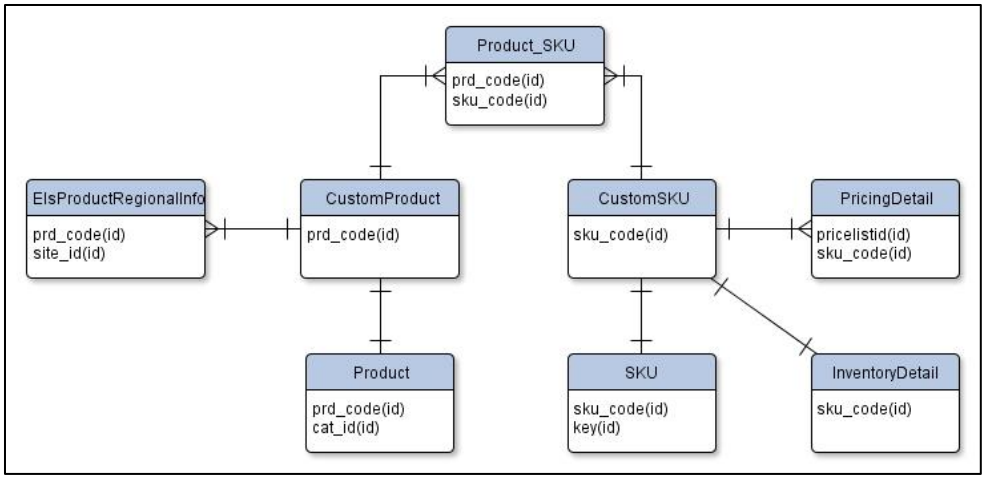

Figure 12 e-commerce system Entities

.

5 Data Quality (DQ) in Elsevier e-commerce

5.1 Data Quality Definition

The interview document for business is using the data quality dimensions that are defined by Morbey **[3]** (Table 15). While the interview document for IS is using definition by Zang (Table 16). Both are selected because they provided a set of dimensions and definition from practitioner's perspective. The introduction of the data quality definition at this phase is important to give the common understanding among interviewees/ parties and to ensure the consistency of definition between phases in the process model. Zang definition can also be found in several practitioner documents for example Platon and IBM.

The development of DQ metrics are closely related with the DQ dimensions/ attributes. Since each data quality study only focus on several dimensions or measurements, it is important that we could also map the data quality dimensions developed by Morbey **[3]** with the dimensions developed by other researchers as in Table 5.

Table 5 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 incl. currency	Accuracy	Accuracy by (in)Validity	Accuracy (Semantic correctness, Syntactic correctness, Precision Factor)	Accuracy	Accuracy, Timeliness
5	Absence of repetitions (free of duplicates)	-	-	-	Uniqueness	Duplication
6	Business referential integrity (integrity)	-	Integrity	-	Referential Integrity	Integrity
7	Completeness (Cross check sums, vertical completeness)	Consistency	Consistency for Integrity	-	-	Consistent
8	Normative consistency	Consistency	Consistency	-	Consistency	Consistency

5.2 Activities for maintaining Data Quality

The need of high product data quality is considered important in e-commerce. It is needed to provide sufficient and correct information about the book or journal for 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 an e-commerce system in Elsevier as follows:

i. Data Quality Check at Data Import [Preventive]

Data Import in e-commerce system conducts the data quality check activities. Product data that do not meet the rules will be reported back to the source system (Figure 3, Data Import). Constraints and referential integrity rules within DBMS are also part of activity to maintain data quality. This activity maintains the data quality with these dimensions:

- Duplication
  This is to ensure the process only add new data or update changed attributes.
- Integrity
  This is to check whether an entity has a referenced record(s). Currently the data import module is checking the referential integrity of the Product and SKU entity.
- Timeliness
  Daily process to update the product catalogue only takes minutes to complete because it is an incremental process.

ii. Manual data quality check by internal staff [Corrective]

The dedicated data steward for the product data in the e-commerce system is absence. The Marketing or Sales staff could check the product information about the website and make changes in the repository when data defect is found. They are using e-commerce system's data update component to make the changes (Figure 3, e-commerce system-Product Availability). This activity maintains the data quality with these dimensions: Completeness, Validity, Integrity, Duplication, Consistency, Accuracy, and Timeliness.

iii. Report from website visitor [Corrective]

Web site visitors could report to Elsevier if they found a data defect in the web page.

5.3 Known DQ problem by Information System (IS)

Known challenges in IS related with data quality are:

i. Data consistency and accuracy because there are several data sources. Related challenge is to have the same product information on several websites

ii. Taxonomy translation from data source to target systems (>20)

iii. Allowing the business user to update the data to maintain the data quality. While it could provide better data quality for individual record in timely manner, there are several issues:

- Fixed data could be reversed to prior value at subsequent data update from data source.
- The update could be still incorrect because business users are not knowledge user for product data.

5.4 Business Problem

5.4.1 List of Business Problems

There are several problems experienced by the business caused by poor product data in e-commerce systems as follows:

i. The Customer does not buy a product due to incomplete information in Book data like description and cover. The impact is potential revenue loss.

ii. The Customer could not browse the site conveniently due to ambiguous data in Book data (subjects). This is also because the need of mapping taxonomy in data sources with hierarchy in the websites.

iii. Unable to run a marketing campaign using AdWords and Email channel because of incomplete information like price, title, and cover.

iv. Incomplete information about product data affects the possibility of internet user to find the data using a search engine. The impact is potential revenue loss because internet users do not find Elsevier Product at the top result within a search engine.

v. There is a possibility of offering unavailable product because of:
   - Inaccurate data from CRM and Journal database
   - Inconsistent data of Journal database and e-commerce system. The case is a salesperson deletes a journal that has been sold to another party in e-commerce system but not in Journal database. The subsequent data update from Journal database will eliminate the changes in e-commerce system.
   - Inaccurate data in e-commerce system, where the product is considered as available while it's not.

   The impact is customer dissatisfaction, unrecognized revenue, ineffective marketing, and potential revenue loss. A case for this business problem is the marketing of unavailable product and results in missed potential revenue of 500k USD.

vi. There is a possibility that some products are not included within a marketing campaign (subject oriented) because of wrongly mapped into a category.

5.4.2 Validation of Business Problems

Currently the report on how much the business problem affects the Elsevier business, i.e. revenue lost or financial cost is not available. Thus, we need to validate the business problems using these following activities:

a. Mapping with KPI

The mapping of the above problems with developed e-Commerce KPI is as in Table 6. We could map each business problem to one or more KPIs.

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)

b. Additional reference

Within the Molla&Licker [3][3] E-commerce System Success Model, the content quality and system quality are the key factors to provide customer satisfaction that lead into a purchase order. Some attributes for the content quality are accuracy, currency, and completeness. Flanagin, et al. [1] provided that the up-to-datedness and completeness of information are among the top three 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 into 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.

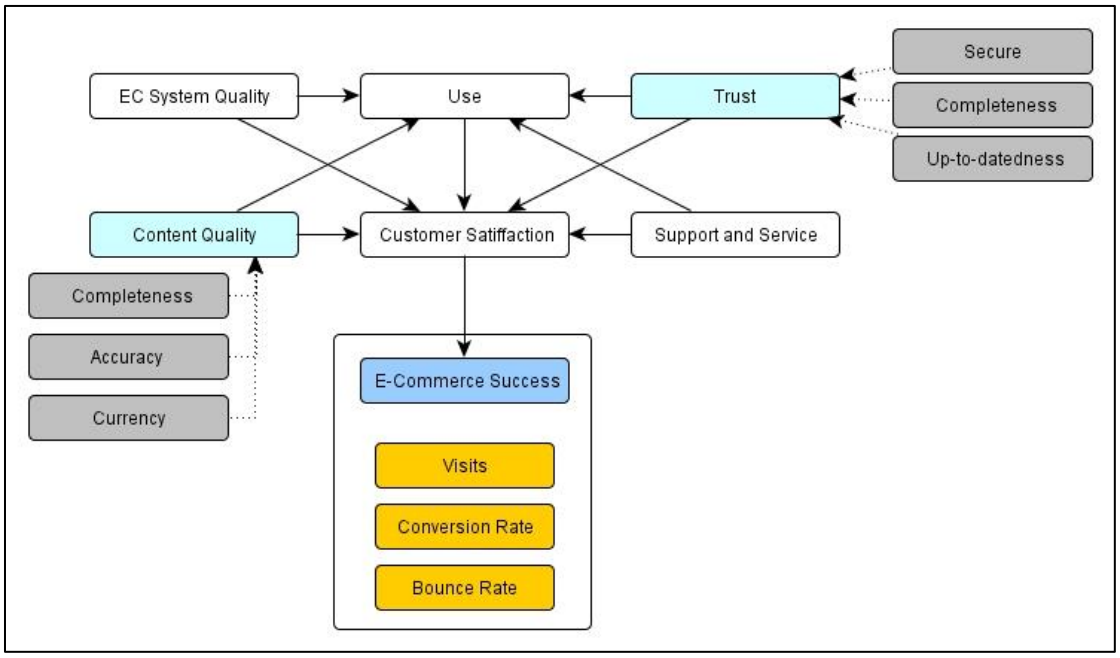

Figure 13 Information Quality and e-commerce, adapted from Molla & Licker [3] Flanagin, et al. [1], Clavis

c. Data Assessment in e-commerce

2

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

An assessment using Google Analytics data is also conducted to support the developed business problems. Data in Google Analytics provides some information about e-store performance in 6 months (08/2013 - 01/2014) period as described in Table 7:

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 expected in 2012 (USD 96.04)
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%) Clickthorugh rates (CTR) are imprints and titles. ▪ A term in Author could **lower** the CTR of Title if combined into 9%.

The assessment result 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 leaving a certain page without further actions. As explained before, the information or content quality is a key factor in e-commerce performance. The low quality of information could lead into frustrating experience, low conversion rate, and lost sales opportunity (Clavis). Some factors related to the product information are the completeness of information, the trust of the customers, and the easiness to find the product from the search engines (Tsai, et al., [6]).

A simple assessment using a Data Quality Tool on the Marketing Dataset (Table 8) supports that poor data quality contributes to the e-commerce performance. The result reveals that some data in e-store is incomplete, incorrect, and inaccurate. This marketing data is derived from the same repository that is used for the e-store Web and it is used as the data source for Google AdWords marketing channel.

Table 8 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

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

No	Defect	Items Affected			Dimensions
		Records	% Records	USD	
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
j	Google Merchant Void	10,801	29.61	1,714,697.15	

*54 of them do not have price in EUR

5.5 Business Problem - Data Defect

Using the previous information about Business Problems and data quality challenges faced by IS we could develop a cause effect matrix as follow:

Context:

The company sells the products through several web based channels. The e-commerce site within this context is e-store (store.elsevier.com) where the product data is managed in a single repository. The e-commerce website sells books and journal in print and electronic format. 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 lower in cost; better customer satisfaction through better accessibility, speed, and higher visibility; and agility in business to adjust to customer's needs. Those are also the challenges in e-commerce. The poor quality of product information could deter the performance expectation.

Table 9 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 e-commerce system (Book)	Completeness per row	All in websites data model
ii	Customer could not browse the site conveniently	Customer dissatisfaction	Ambiguous data in Book (taxonomy mapping problem)	Absence of contradiction	Subject/ Category
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
iv	Internet user could not find the data in top result using search engine	Potential revenue loss	Incomplete information in e-commerce system	Completeness per row	All in websites data model
v	Offering unavailable product	Customer dissatisfaction, unrecognized revenue, ineffective marketing, and potential revenue	Inaccurate data in e-commerce system (Journal)	Accuracy inc. currency	Saleable/ Availability in a Region
			Incomplete data in e-commerce system (Journal)	Completeness, Business Referential Integrity	Fulfillment system

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

Table 10 Preventive and Reactive Measures

No	Type	Type	DQ Dimensions	Attribute
i	DQ Check at Data Import	Preventive	Currency, Business Referential Integrity, Absence of Repetition	All in repository data model
ii	Manual update using e-commerce system	Reactive	Completeness per row, Business Referential Integrity, Accuracy incl. currency Absence of contradictions, Absence of repetitions,	All in website data model

Appendix 1 Interview Documents

1 Preliminary Document for e-commerce Business
2 Interview with e-commerce Business
3 Preliminary Document for e-commerce System
4 Interview with e-commerce System

Appendix 2 e-commerce Metrics

Table 11 e-commerce KPI, Tsai, et al. [5]

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 3 Data Model for Elsevier e-commerce Web sites - Comparison

LN = LexisNexis

HS = HealthScience

1. Journal

Table 12 Journal Information in Elsevier and competitors e-commerce site

e-store Elsevier	LN	HS	myElsevier	Wiley General	Wiley Detail	Springer	SAGE	Description (e-store Elsevier)
Journal Info								
Title	v	v	v	v	v	v	v	The data is retrieved from promis25810.htm
Subtitle	x	x	x	x	x	v	v	The data is retrieved from promis25810.htm
Format	v	x	x	v	v	v	v	Journal if the type is ISSN
Impact Factor	x	v	v	v	v	v	v	This data is retrieved from pr_impactfactor.csv
Impact Factor Year	x	v	v	v	v	v	v	This data is retrieved from promis25810.htm
5 Yr Impact Factor	x	x	x	x	x	x	x	This data is retrieved from 5yearimpactfactor.csv
Editor Information	x	v	x	v	v	v	x	This data is retrieved from pr_edboard.txt
Popup for Editors	x	x	x	x	x	x	x	This data is retrieved from pr_edboard.txt
ISSN	v	v	v	v	v	v	v	This data is retrieved from promis25810.htm
Journal Link	x	x	x	v	x	x	x	This is the linking to Elsevier.com based on ISSN
Volume Number	x	x	x	v	v	v	v	This data is retrieved from promis25810.htm
Issue Information	x	x	x	x	x	x	x	This data is retrieved from promis25810.htm
short description	v	v	v	v	v	v	v	This data is retrieved from promis25810.htm
Cover	v	v	v	v	v	v	v	
Overview								
Aims & Scope	x	x	x	x	x	v	v	This is the proddesc data retrieved from promis25810.htm
Audience	x	x	x	x	x	x	x	this is the col014 data retrieved from promis25810.htm
Editorial Information								
Editor Information	x	x	x	x	v	v	x	

e-store Elsevier	LN	HS	myElsevier	Wiley General	Wiley Detail	Springer	SAGE	Description (e-store Elsevier)
Bibliographic								
Dispatch Dates	x	x	x	x	x	x	x	This is the dispatch dates. This data is fetched from screen scraping of Elsevier.com
Abstracting and Indexing	x	x	x	x	x	x	x	This is the abstract information. The data is retrieved from promis25810.htm
Issue								
Free Sample link	x	x	x	x	v	v	x	Free sample issue flag is based on data from pr_relatedurls_t2.csv file
Special Issues	x	x	x	v	v	v	v	This data is fetched from screen scraping of Elsevier.com
Articles								
Articles	v	x	x	x	v	v	v	This data is fetched from screen scraping of Elsevier.com
Price								
Format	v	x	x	x	v	x	v	
Price	v	v	v	x	x	x	v	Prices for Journals are from two sources a. CRM – eJournals b. PROMIS – Print Journals
Stock Information	v	x	x	x	x	x	x	

2. Book

Table 13 Book Information in Elsevier and competitors e-commerce site

e-store Elsevier	HS	LN	MyElsevier	Amazon	Barnes and Nobles	Description (e-store Elsevier)
Book Info						
Title	v	v	v	v	v	NOTE: Title on all pages except product detail page is a combination of 3 values. Title, Subtitle and Edition. (i.e. search results list)
Edition	v	v	x	v	v	
Sub Title	v	x	x	x	x	
Product type Map	x	v	x	v	v	The products are at high level classified as Print Book, eBook and Journal
Name of the Author	v	v	x	v	v	Currently last names of the author is displayed with links to the Author Page
Publication Date	v	v	x	v	v	
Imprint	v	v	x	v	v	.
The next edition	v	x	x	x	x	The Value is a product in the PPM relationship table with a relationship ID of 1 (Edition)
Printbook ISBN	v	v	v	v	v	ISBN 13
eBook ISBN	x	v		x	v	
Short Description	x	x	x	x	x	PPM Field: ExtB-Mkt/Sls – Tease (for ALL markets)
Series	x	x	x	v	v	If the product belongs to Series, the series is linked
Pages	v	x	x	v	v	This is the page count value
Dimension	v	x	x	v	v	Trim Height and Trim width
Cover	v	v	v	v	v	
Overview						
Key Feature	v	x	x	x	x	PPM Promo Value Field for ALL markets - ExtB Key Features
Long Description	v	v	v	v	v	PPM Promo Value for ALL markets – ExtS-General Description
Readership	x	x	x	x	x	ExtS-Mkt/Sales - Audience
Bread Crumb	x	x	x	x	x	this breadcrumb is created dynamically by utilizing category hierarchy
Authors						

e-store Elsevier	HS	LN	MyElsevier	Amazon	Barnes and Nobles	Description (e-store Elsevier)
Display Name and biography	v	v	x	v	v	
Affiliations & Expertise	v	v	x	v	v	PPM Field under Authors: Notes Type: Affiliations
Recent Publicaton	x	x	x	v	v	4 titles published by the author sorted by publication date
TOC						
Title	x	x	x	x	x	Title of the product
TOC	v	v	x	x	v	ExtB-TOC – Long
Video						
Video	x	x	x	x	x	Media File(s)
Learn More						
Online Companion Materials link	x	x	x	x	x	
Instructor Ancillary Support Materials link	x	x	x	x	x	
Editorial Review						
Review	x	x	x	v	v	The section is populated from the PPM Field: ExtS Mkt/SlsQuotes
Price						
Format	v	v	x	v	v	
Price	v	v	v	v	v	*Print Books and non Science Direct e-Books* – PPM *Science Direct, eSubscription e-Books*– CRM
Stock Information	v	v	x	v	v	

Appendix 4 Data Model for Elsevier e-commerce Database

Table 14 Data Model in ATG Database

ATG Data		Used#	Journal		Book					eBook			
Type	**Fields**		**UK**	**US**	**ANZ**	**APAC**	**UK**	**US**	**Tango 2**	**DE**	**FR**	**UK**	**US**
Product	Prd_Code	11	v	v	v	v	v	v	v	v	v	v	v
	Prd_Display_Name	11	v	v	v	v	v	v	v	v	v	v	v
	Prd_Description	11	v	v	v	v	v	v	v	v	v	v	v
	Prd_Long_Description	11	v	v	v	v	v	v	v	v	v	v	v
CustomProduct	ABSTRACT	7	v	v	v	v	v	v	v				
	AdditionalInfo	6	v	v	v	v	v	v					
	AllAuthorDesc	11	v	v	v	v	v	v	v	v	v	v	v
	ANCILLARYTITLE	6	v	v	v	v	v	v					
	ApprovalRequired	6	v	v	v	v	v	v					
	AudienceType	10	v	v	v	v	v	v		v	v	v	v
	Author	6	v	v	v	v	v	v					
	AUTHORALIST	11	v	v	v	v	v	v	v	v	v	v	v
	AUTHORBLIST	10	v	v	v	v	v	v		v	v	v	v
	AvailableDate	10	v	v	v	v	v	v		v	v	v	v
	Benefits	6	v	v	v	v	v	v					
	Brochure	6	v	v	v	v	v	v					
	Case_Studies	6	v	v	v	v	v	v					
	Comments	7	v	v	v	v	v	v	v				
	COMMODITYCODE1	10	v		v	v	v	v	v	v	v	v	v
	COMMODITYCODE2	6	v		v	v	v	v	v				
	Compcopy	6	v	v	v	v	v	v					
	CONTRACTORORIGIN	11	v	v	v	v	v	v	v	v	v	v	v
	COPSPRODUCTTYPEID	11	v	v	v	v	v	v	v	v	v	v	v
	COPYKEYFEATURE	10	v	v	v	v	v	v		v	v	v	v
	COPYRIGHT	11	v	v	v	v	v	v	v	v	v	v	v
	CreatedByTitle	4		v	v	v		v					
	CurrencyType	6	v	v	v	v	v	v					
	DetailDesc1	7	v	v	v	v	v	v	v				
	EditionNumber	11	v	v	v	v	v	v	v	v	v	v	v
	EditionText	11	v	v	v	v	v	v	v	v	v	v	v
	EDITOR	7	v	v	v	v	v	v	v				
	ELEMENTPERCENTAGE	9	v		v	v	v	v		v	v	v	v
	EnableE-commerce	11	v	v	v	v	v	v	v	v	v	v	v
	EXTNEWTOEDITION	10	v	v	v	v	v	v		v	v	v	v

ATG Data		Used#	Journal		Book					eBook			
Type	**Fields**		**UK**	**US**	**ANZ**	**APAC**	**UK**	**US**	**Tango 2**	**DE**	**FR**	**UK**	**US**
	EXTORGCOPY	6	v	v	v	v	v	v					
	FREQUENCY	7	v	v	v	v	v	v	v				
	FULFILLMENTCOMPANY CODE	10	v	v	v	v	v	v		v	v	v	v
	FullRegistration	0											
	HOSTED	0											
	IMPRINT	11	v	v	v	v	v	v	v	v	v	v	v
	IsActive	11	v	v	v	v	v	v	v	v	v	v	v
	Isbn	11	v	v	v	v	v	v	v	v	v	v	v
	Isbn10	7	v	v	v	v	v	v	v				
	ISBN10FORMATTED	11	v	v	v	v	v	v	v	v	v	v	v
	ISBN13FORMATTED	7	v	v	v	v	v	v	v				
	IsTemplate	6	v	v	v	v	v	v					
	Keywords	6	v	v	v	v	v	v					
	LINK	1							v				
	MrktUrl	6	v	v	v	v	v	v					
	NEXTEDITIONISBN	7	v	v	v	v	v	v	v				
	NUMBEROFILLUSTRATIONS	0											
	NUMBEROFPAGES	11	v	v	v	v	v	v	v	v	v	v	v
	PARENTTITLE	11	v	v	v	v	v	v	v	v	v	v	v
	PeerReview	6	v	v	v	v	v	v					
	Platform	11	v	v	v	v	v	v	v	v	v	v	v
	PLATFORMREQUIREMENT	7	v	v	v	v	v	v	v				
	Prd_Code	11	v	v	v	v	v	v	v	v	v	v	v
	PREVISBN	7	v	v	v	v	v	v	v				
	PriceText	6	v	v	v	v	v	v					
	PricingDescription	6	v	v	v	v	v	v					
	PRODUCTCLASSID	7	v	v	v	v	v	v	v				
	ProductFormat	10	v	v	v	v	v	v		v	v	v	v
	PRODUCTLANG	6	v	v	v	v	v	v					
	ProductLine	0											
	ProductType	11	v	v	v	v	v	v	v	v	v	v	v

ATG Data		Used#	Journal		Book					eBook			
Type	**Fields**		**UK**	**US**	**ANZ**	**APAC**	**UK**	**US**	**Tango 2**	**DE**	**FR**	**UK**	**US**
	PUBCODE	6	v	v	v	v	v	v					
	PubDate	10	v	v	v	v	v	v		v	v	v	v
	PUBDATETYPE	6	v	v	v	v	v	v					
	PubNumLog	5			v	v	v	v	v				
	Pubstatus	10	v	v	v	v	v	v		v	v	v	v
	Rank	0											
	REGION	0											
	Reviews	6	v	v	v	v	v	v					
	SDLINK	6	v	v	v	v	v	v					
	Searchable	7	v	v	v	v	v	v	v				
	SERIESID	6	v	v	v	v	v	v					
	SERIESPROMO	7	v	v	v	v	v	v	v				
	ShortDescription	10	v	v	v	v	v	v		v	v	v	v
	Status	7	v	v	v	v	v	v	v				
	STOCKSTATUS	10	v	v	v	v	v	v		v	v	v	v
	SUBTITLE	5			v	v	v	v	v				
	TaxProductCode	11	v	v	v	v	v	v	v	v	v	v	v
	TeaseCopy	7	v	v	v	v	v	v	v				
	Title	11	v	v	v	v	v	v	v	v	v	v	v
	Toc	7	v	v	v	v	v	v	v				
	TRTCODE	7	v	v	v	v	v	v	v				
	UpdatedDate	7	v	v	v	v	v	v	v				
	VendorID	11	v	v	v	v	v	v	v	v	v	v	v
	VOLUMENUMBER	7	v	v	v	v	v	v	v				
	WEBPRODUCTTYPENAME	6	v	v	v	v	v	v					
	Weight	7	v	v	v	v	v	v	v				
	Year	6	v	v	v	v	v	v					
Sku	SkuCode	11	v	v	v	v	v	v	v	v	v	v	v
	SkuDisplayName	11	v	v	v	v	v	v	v	v	v	v	v
	SkuDescription	10	v	v	v	v	v	v		v	v	v	v
	Fulfiller	10	v	v	v	v	v	v		v	v	v	v
CustomSku	Active	11	v	v	v	v	v	v	v	v	v	v	v
	FulfillmentCompanyCode	1							v				
	GUESTEDITOR	7	v	v	v	v	v	v	v				
	Isbn	11	v	v	v	v	v	v	v	v	v	v	v
	Issue	2	v	v									
	MISC1	1							v				

ATG Data		Used#	Journal		Book					eBook			
Type	**Fields**		**UK**	**US**	**ANZ**	**APAC**	**UK**	**US**	**Tango 2**	**DE**	**FR**	**UK**	**US**
	MISC2	1							v				
	MISC3	1							v				
	MONTH	7	v	v	v	v	v	v	v				
	OUTOFPRINT	6	v	v	v	v	v	v					
	PurchaseLimit	1							v				
	PURCHASETYPE	10	v	v	v	v	v	v		v	v	v	v
	SkuCode	11	v	v	v	v	v	v	v	v	v	v	v
	SkuType	10	v	v	v	v	v	v		v	v	v	v
	TOPIC	11	v	v	v	v	v	v	v	v	v	v	v
	VOLUME	1	v										
	YEAR	6	v	v	v	v	v	v					
PricingDetail	SkuCode	10	v	v	v	v	v	v		v	v	v	v
	PricingScheme	11	v	v	v	v	v	v	v	v	v	v	v
	ListPrice	11	v	v	v	v	v	v	v	v	v	v	v
	PriceListId	11	v	v	v	v	v	v	v	v	v	v	v
	Isbn	11	v	v	v	v	v	v	v	v	v	v	v
InventoryDetail	Sku_Code	11	v	v	v	v	v	v	v	v	v	v	v
	Availability_Status	11	v	v	v	v	v	v	v	v	v	v	v
	Availability_Date	11	v	v	v	v	v	v	v	v	v	v	v
ElsProductRegionalInfo	Audience	1							v				
	EnableE-commerce	1							v				
	FulfillmentCompanyCode	1							v				
	isActive	1							v				
	KeyFeature	1							v				
	LongDescription	1							v				
	ProductCode	1							v				
	PubDate	1							v				
	PubStatus	1							v				
	Quotes	1							v				
	ShortDescription	1							v				
	SiteId	1							v				
ElsLinkedFormatsMulti	ProductId	5							v	v	v	v	v
	skutype	1							v				
	skuid	1							v				
parentcategory		10	v	v	v	v	v	v		v	v	v	v

Appendix 5 Appendix 3. Data Quality Dimension

Table 15 Selected Data Qualities, Morbey [3]

No	Dimensions	Definition	Common Dimension
1	Completeness per row (horizontal completeness)	Is there any missing or defective data in a record? All data is entered according to business needs.	Completeness
2	Syntactical correctness (conformity)	Is there data in a non-standardized format? The data fits into the specific format	Validity
3	Absence of contradictions (consistency)	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).	Consistency
4	Accuracy incl. currency and timeliness	Which data is wrong or expired? Correct and up to date (timeliness) notation of existing names, addresses, products etc.	Accuracy Timeliness
5	Absence of repetitions (free of duplicates)	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).	Duplication
6	Business referential integrity (integrity)	Which reference data or relations are missing? There will not be any clients without a contract, products will be listed.	Integrity
7	Completeness (Cross check sums, vertical completeness)	Is there data consistency over all systems? For instance: at an appointed date the number of contracts in the data source is exactly the same as the number of contracts in the DWH.	Validity
8	Availability of documentation (find ability)	Can the data be found easily and quickly (e.g. using common “search”-functions)	
9	Normative consistency	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.	Consistency
10	Quality of Information sources	this dimension is composition of believability, reputation, objectivity, and reliability (credibility) dimensions	Validity

Table 16 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?

References

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

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

[3] 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.

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

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

[6] 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.