Electron Markets (2011) 21:141–154
DOI 10.1007/s12525-011-0059-x

SPECIAL THEME

Product data quality in supply chains: the case of Beiersdorf

Kai M. Hüner · Andreas Schierning · Boris Otto · Hubert Österle

Received: 24 June 2010 / Accepted: 26 January 2011 / Published online: 4 May 2011

Abstract A number of business requirements (e.g. compliance with regulatory and legal provisions, diffusion of global standards, supply chain integration) are forcing consumer goods manufacturers to increase their efforts to provide product data (e.g. product identifiers, dimensions) at business-to-business interfaces timely and accurately. The quality of such data is a critical success factor for efficient and effective cross-company collaboration. If compliance relevant data (e.g. dangerous goods indicators) is missing or false, consumer goods manufacturers risk being fined and see their company's image damaged. Or if logistics data (e.g. product dimensions, gross weight) is inaccurate or provided not in time, business with key account trading partners is endangered. To be able to manage the risk of business critical data defects, companies must be able to a) identify such data defects, and b) specify and use metrics that allow to monitor the data's quality. As scientific research on both these issues has come up with only few results so far, this case study explores the process of identifying business critical product data defects at German consumer goods manufacturing company Beiersdorf AG. Despite advanced data quality management structures such defects still occur and can result in complaints, service level impairment and avoidable costs. The case study analyzes product data use and maintenance in Beiersdorf's ecosystem, identifies typical product data defects, and proposes a set of data quality metrics for monitoring those defects.

Keywords Case study · Data quality · Data quality management · Data quality metrics · Product data · Supply chain management

JEL L15 · L66

Responsible editor: Hans-Dieter Zimmermann

K. M. Hüner (✉) · B. Otto · H. Österle
University of St. Gallen, Institute of Information Management,
Müller-Friedberg-Strasse 8,
9000 St. Gallen, Switzerland
e-mail: kai.huener@unisg.ch

B. Otto
e-mail: boris.otto@unisg.ch

H. Österle
e-mail: hubert.oesterle@unisg.ch

A. Schierning
Beiersdorf AG, Supply Chain Data Process Management,
Unnastraße 48,
20245 Hamburg, Germany
e-mail: andreas.schierning@beiersdorf.com

Introduction

A number of business requirements (e.g. compliance with regulatory and legal provisions, the ongoing consolidation process in the retail industry, diffusion of global standards) are forcing consumer goods manufacturers to increase their efforts to provide product data (e.g. identifiers, dimensions) at business-to-business interfaces timely and accurately. The following examples are supposed to illustrate the situation.

Large retailers such as Walmart (approx. 422 billion USD revenue in 2010), Carrefour (approx. 90 billion EUR revenue in 2010) or Metro (approx. 67 billion EUR revenue in 20010) in recent years have put up strict requirements on cross-company processes, which consumer goods manufacturers are expected to meet.

Carrefour, for example, demands that logistics related product data be provided as early as eight months before a product's launch in order to ensure high-quality intra-company supply chain planning. In order to adequately respond to this challenge, consumer goods manufacturers in general need to enhance their innovation and marketing processes by an integrated product data management process allowing to capture relevant data at an early stage and provide this data timely and in good quality.

Global Standards One (GS1), a leading standardization organization, has specified the Global Trade Item Number (GTIN), allowing to unambiguously identify trade items all over the world (GS1 2010, pp. 22–77). Based on this numerical system, external service providers offer data pools for multilateral synchronization of product data (Nakatani et al. 2006, pp. 970–977). Consumer goods manufacturers making their product data available in such a pool are required to label every traded item so that each item and each logistic unit (e.g. shrink, packaging unit, pallet) is unambiguously identifiable by a unique GTIN.

The present paper primarily aims at a) identifying typical business critical product data defects occurring in a consumer goods manufacturer's ecosystem and b) proposing a set of data quality metrics for monitoring those defects. As product data defects and their business impact have been rarely investigated in scientific studies, case study research constitutes an appropriate research methodology. The paper is structured in five sections and continues with a brief overview of the state of the art regarding data quality measuring and cross-company master data management, followed by a methodological justification of the presented case study. The next section describes the initial situation at Beiersdorf and reports on how business critical defects were identified and how data quality metrics were specified. Identified defects and specified metrics are presented as well.

Finally, the paper concludes with a short summary of the results and a brief discussion of implications for practice and research.

Data quality in cross-company supply chains

Data quality management

Various studies conceptualize data quality by data quality dimensions (e.g. accuracy, completeness, objectivity) on the basis of empirical research (Wang and Strong 1996), ontological and semiotic conclusions (Price and Shanks 2005; Wand and Wang 1996) or practitioners' experiences (English 1999, pp. 87–118; Redman 1996, pp. 245–266). These studies have established the definition of data quality as the data's 'fitness for use', i.e. whether data is of good or poor quality depends on the context it is used in and the user it is used by. This context dependent data quality definition provides the basis for the specification of business-oriented data quality metrics: Data is of good quality if it meets business requirements (e.g. requirements of business partners or internal processes). Of course, requirements regarding one data object may vary from one process to another.

Data quality management comprises activities for improvement of data quality (Batini and Scannapieco 2006, pp. 69–71). Going beyond mere reactive action (e.g. identification and correction of data defects) (Shankaranarayanan and Cai 2006, pp. 303–304), data quality management works as a preventive concept, characterized by a continuous cycle consisting of activities to define, measure, analyze and improve data quality (English 1999, pp. 69–81; Eppler and Helfert 2004; Wang et al. 1998). Preventive data quality management includes the design and deployment of appropriate management structures such as data governance (Khatri and Brown 2010; Weber et al. 2009) or the specification and implementation of data quality metrics (Heinrich et al. 2009). An overview of the most relevant approaches for data quality management is given by Batini et al. (2009).

Data quality measurement

Apart from the conceptualization of data quality by means of various data quality dimensions the measurement of data quality (i.e. of data quality dimensions) has been a central issue of many scientific studies. In this context, data quality metrics serve to operationalize data quality dimensions: They specify the data to be measured, a measuring point, a measuring technique, and a measuring scale (Batini and Scannapieco 2006, p. 19). Some studies provide calculation instructions for single data quality dimensions, such as timeliness (Heinrich and Klier 2009; Heinrich et al. 2007), accuracy (Gorla and Krishnan 2002), or completeness (Cai and Ziad 2003). Other studies present procedures and techniques for measuring data quality by means of interviews and surveys (Lee et al. 2002; Nicolaou and McKnight 2006; Price et al. 2008) or by means of validation rules (Fan et al. 2008; Hipp et al. 2007). For the process of identifying data defects and specifying data quality metrics in a specific context procedure models and analysis techniques have been proposed as well (Batini et al. 2007; Caballero et al. 2008; Gelinas and Dull 2009; Heinrich and Klier 2009). For the design of business-oriented data quality metrics (i.e. metrics for monitoring business critical data defects) causal relations between data defects, business operations problems, and strategic business goals should be analyzed (Otto et al. 2009). The objective is to identify those process activities a) the outcome of which is critical in achieving strategic business goals, and b) the outcome of which is strongly dependent on the quality of the data used. Examples of such causal relations are presented in following sections.

Data sharing in supply chains

Studies that investigate cross-company data exchange distinguish between exchange of transaction data (for example, wholesalers and retailers reporting sales figures to the manufacturer to facilitate demand planning) and exchange of master data (Legner and Schemm 2008, p. 126). While the former has been investigated in numerous scientific studies (Cachon and Lairiviere 2000; Chen et al. 2000; Kulp et al. 2004; Lau et al. 2002; Lee et al. 1997; Sheu et al. 2006), the latter has somewhat been neglected. Although it is common sense that electronic data exchange in supply chains (Beck and Weitzel 2005) and electronic marketplaces (Wang et al. 2006) hold benefits for most companies, and that correct interpretation of exchanged data is business critical (Goh et al. 1999; Madnick 1995; Vermeer 2000; Zhao 2007), business impacts of the quality of exchanged master data (both intra-company and cross-company) have not been thoroughly examined so far. Therefore, this paper investigates business problems that are caused by product master data defects. For monitoring those defects, the paper proposes data quality metrics that have been designed at Beiersdorf.

In general, master data specifies business objects, i.e. those essential business entities a company's business activities are based on. Such entities are, for example, business partners (customers, suppliers), products and the components and materials of which they are comprised, or employees (Smith and McKeen 2008, pp. 64–65). Basically, master data can be differentiated by three concepts: master data class, master data attribute, and master data object (Loshin 2008, pp. 5–8). A master data object represents a concrete business object (a product manufactured in a certain plant at a certain point in time, for example), and it specifies selected characteristics of this business object (color, features, or price, for example) by means of attributes. The attributes selected for representation of a specific class of business objects (customers or products, for example) constitute a master data class (which is usually specified by a data model).

Research methodology

The underlying research method of the presented work is participative exploratory case research (Scholz and Tietje 2002, pp. 11–12; Yin 2002, pp. 3–5). The scientific discipline of information systems research often uses case study research to deal with organizational issues in which the boundaries between the real-world phenomenon itself (here: product data quality at the interfaces of Beiersdorf and their business partners) and its environment (here: the ecosystem of Beiersdorf's supply chain) cannot clearly be identified (Yin 2002, p. 13).

The present case study aims at

- describing a method based process of identifying business critical data defects,
- analyzing product data use and maintenance in a consumer goods manufacturer's ecosystem in order to identify typical product data defects, and
- proposing a set of data quality metrics for monitoring product data.

Conducting research this way is reasonable if a phenomenon has not been investigated to a large extent so far, or if a phenomenon has not been fully understood theoretically, or if during a research process questions to be tackled by future research are raised (van der Blonk 2003; Walsham 1995).

The data for this case study was collected over the period of four months (from February 2010 to June 2010, see Table 1) from varied sources including various

Table 1 Project meetings and interviews

Date	Organizational unit	Meeting type (and topic)	Number of participants	Duration (hours)
February 15, 2010	Data Process Management	Project meeting (project planning)	3	3
February 26, 2010	Data Process Management	Project meeting (attribute selection and grouping, interview preparation)	3	6
March 3, 2010	Quality Management (Distribution)	On-site interview in person	1	2
March 17, 2010	Distribution Center (Eastern Europe)	Invited group interview in person	3	4
March 24, 2010	Third Party Manufacturer (mass aerosol products)	On-site group interview in person	2	3
March 30, 2010	Marketing and Sales	On-site group interview in person	2	3
March 31, 2010	Distribution Center (Central Europe)	On-site group interview in person	2	3
April 1, 2010	Plant (complex care products, *Eucerin*)	Invited group interview in person	2	3
April 7, 2010	Plant (mass products, *Nivea*)	On-site interview in person	4	4
April 26, 2010	Data Process Management	Project meeting (interview analysis)	3	6
May 10, 2010	Data Process Management	Project meeting (metrics specification)	3	6
June 16, 2010	Data Process Management, Beiersdorf Shared Services, External Software Provider	Project meeting (planning of metrics implementation)	6	4

company documents, observations made from attending project meetings and seven semi-structured interviews. This multi-method approach can help to generate data that is often rich in detail and rigor (Yin 2002, pp. 83–101). As the authors of this paper have contributed methodology and concepts (i.e. the procedure of identifying data defects and specifying data quality metrics) to the observed project, presented results might contain biased interpretation as it is due to participative case studies (Baskerville 1997, p. 29). However, during the course of interviews (i.e. the process of data collection), the authors have not been involved in discussions or the process of identifying data defects. Interview results (i.e. issue descriptions, ratings) were recorded in writing and were reviewed and approved by interviewees.

The case of Beiersdorf

Current situation

Company profile

Beiersdorf, with headquarters in Hamburg, Germany, is a global company having more than 150 subsidiaries and employing over 19,000 people all over the world (Beiersdorf 2010). The company is divided in two business segments. The Consumer business segment (approx. 5.3 billion EUR revenue in 2010) offers products for the skin and beauty care market as well as medical products. *Nivea*, one of the oldest and most popular body care brands in the world, comprises a set of products contributing the biggest portion to the company's revenue. Other well-known products in this field are 8×4 and *Labello* in the mass distribution market, and *Juvena* and *La Prairie* in the exclusive-products market. Medical products offered comprise plasters, adhesive tape, sore protection, and bandages, distributed under brands such as *Eucerin* and *Hansaplast*. The *tesa* business segment (approx. 873 million EUR revenue in 2009) comprises self-adhesive system and product solutions both for the industry and for the consumer market.

Beiersdorf's organizational structure is divided into two functional and three regional areas of responsibility (see Fig. 1). Information technology is managed by Beiersdorf Shared Services, a shared service organization and subsidiary of Beiersdorf. The organizational unit Supply Chain Data Process Management, located in the corporate supply chain technology division, is responsible for organization of company-wide product master data management.

Ecosystem

As the focus of the case study is on cross-company data exchange, Fig. 2 shows a segment of Beiersdorf's ecosystem in which corporate functions form only one of eleven actors. The illustration differentiates between intra-company data flows (flows 1, 2, 8, and 12), data flows between Beiersdorf and external parties (flows 3, 4, 5, 6, 7, 9, 10, 11, 13, and 14) and flows between non-Beiersdorf entities (flows 15 and 16). Application systems involved in the exchange of data are not shown in the figure.

In the focus of the case study are the following actors and data flows from Beiersdorf's ecosystem:

- *Corporate function*. Central company functions (e.g. demand planning, product marketing, product development, packaging design) use and maintain product data (see [1] in Fig. 2; for example, a product recipe is created by the product development department and used by the demand planning function) and make product data available (for example, GTINs or net weights included in an artwork specification [4]) for external partners).
- *Production plant, supplier, third party manufacturer, artwork agency*. Globally distributed production plants and external production partners use, for example, item lists or GTINs [2, 7] as well as country specific

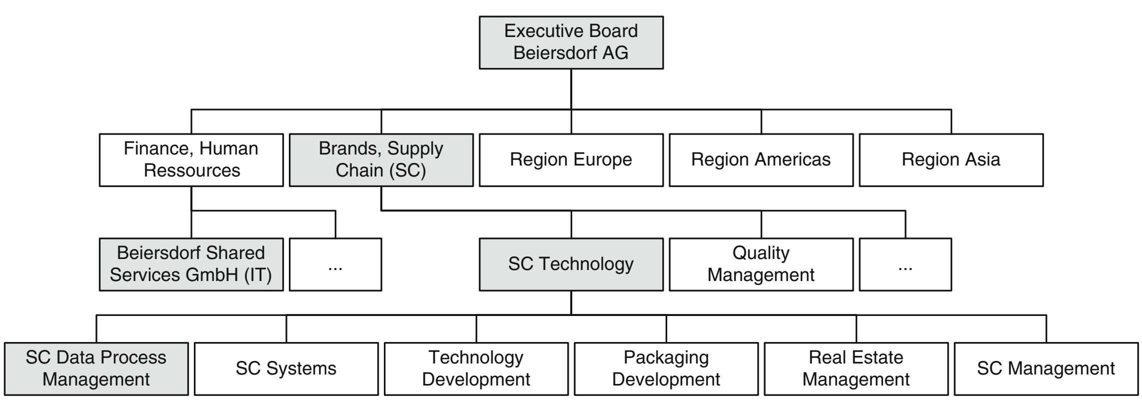


Fig. 1 Organizational structure of Beiersdorf and reporting line of Data Process Management

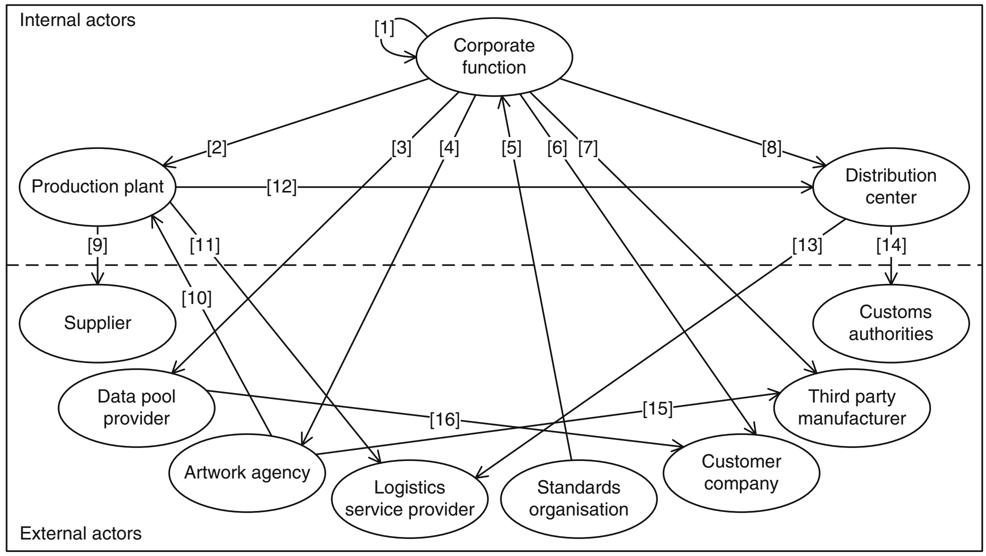


Fig. 2 Product data exchange in the ecosystem of Beiersdorf

artworks [10, 15], and order raw materials (positions of bills of material [9]) from suppliers.

- *Distribution center, logistics service provider, custom authorities*. For storage and transportation of goods, distribution centers use logistics data provided by corporate functions [8] and modified or added by production plants [12]. External service providers also use logistics data as well as environment, health and safety data [11, 13]. Customs relevant data [14] need to be made available to customs authorities.
- *Data pool provider*. Beiersdorf makes GTINs [3] available in a data pool to be used by their customer companies [16].
- *Customer company*. Beiersdorf provides logistics data to customer companies in order to support their planning processes ([6], particularly data on packaging dimensions).
- *Standards organization*. Beiersdorf requests data from standardization organizations for definition of new GTINs [5].

Data management organization

Tasks to be done by the Data Process Management unit (see Fig. 1) comprise typical duties of a Chief Data Steward (Weber et al. 2009, pp. 10–13), such as strategic development of master data management or further development of the central master data system. System support is provided by Beiersdorf Shared Services.

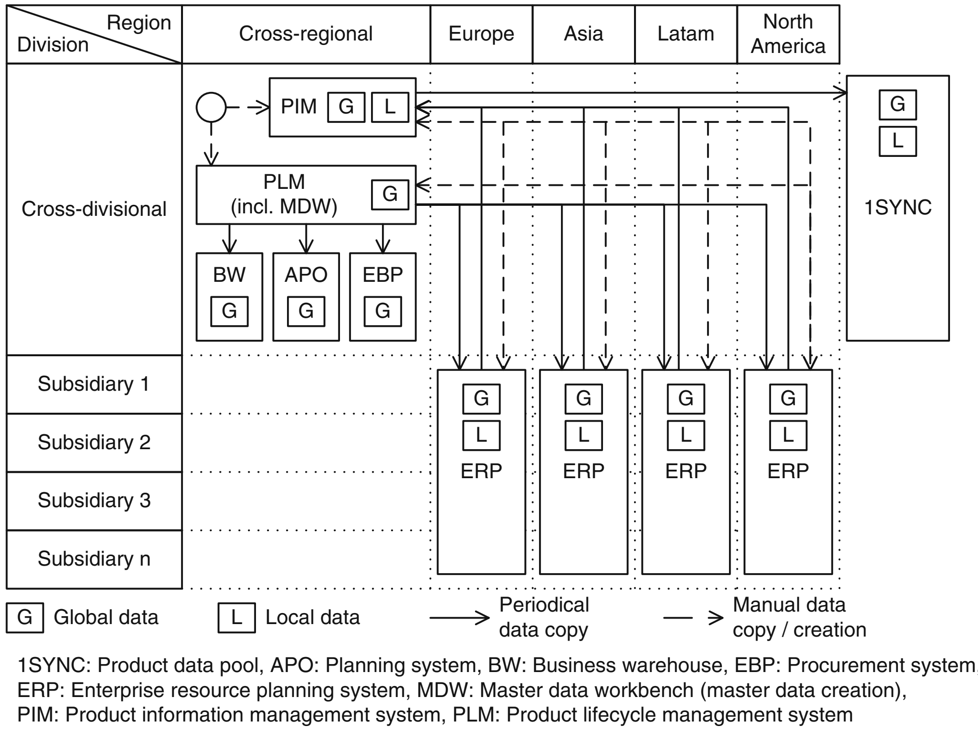
The head of Supply Chain (see Fig. 1) represents executive sponsorship for master data management at Beiersdorf. Counterparts of Data Process Management are people responsible for master data management in the different corporate or local functional departments, who take over a coordinating function for all master data related issues. This role is fulfilled by various central or local organizational units. On a corporate level, one person from the marketing department is chosen to be responsible for each product line (e.g. *Nivea Sun*, *Nivea Body*). And on the level of subsidiaries usually one Business Data Steward from the material management function is appointed per country.

Responsibility for the creation of the master data object of a specific product depends on the market the product is supposed to be sold. Product data representing products marketed only in one country is created by the local subsidiaries therein, whereas product data of internationally distributed products is created by brands on a corporate level. Who is then responsible for further maintenance depends on where the data was created. For the central company functions approx. 200 people are involved in the process of maintaining global data attributes, whereas on the local level usually the product management function manages master data (Schemm 2008, pp. 230–232).

Data management systems

For the management of global product data, such as identifiers, logistics data, and product hierarchies, Beiersdorf is using a central product lifecycle management (PLM) system, which was implemented in 2004. The PLM system at regular intervals (i.e. every 3 hours) provides new or changed product data to four regional enterprise resource planning (ERP) systems and a number of other global application systems (e.g. a decision support system (BW), a planning system (APO), a procurement system (EBP)). As the data is directly committed into the receiving systems, a consistent database is always ensured. The systems are operated by Beiersdorf Shared Services. Figure 3 illustrates the flows of master data within Beiersdorf's application landscape. The application architecture depicted is typical for a global company, comprising both global applications

Fig. 3 Master data flows within the application landscape of Beiersdorf



supporting processes that affect several organizational units and local applications supporting processes within discrete organizational units (Lehmann 2003).

Being part of the PLM system, the Master Data Workbench (MDW) provides functionality for master data creation, thereby ensuring that master data is captured by the PLM system right at the moment it is being created. The users of the system (about 150) work with input masks specifically designed to match with the product innovation process and the product data maintenance process. The whole process of master data gathering and creation is done by means of one single application system. Fast and accurate data processing is ensured, since there is no media breakage and a number of PLM functions (e.g. allocation of unique identifiers, check routines) can already be used in the process of data creation. After the release of a product master data record, the PLM system provides product data to the regional ERP systems.

Product data distribution to wholesalers and retailers is controlled by means of a central product information management (PIM) system. The PIM system is provided with global and local master data from the regional ERP systems. The PIM system also controls the data transfer to 1Sync, Beiersdorf's data pool.

Data quality measurement

Project objectives

Basically, the given organizational and technical master data management structures at Beiersdorf have been allowed for smooth cross-company supply chain processes. Regularly occurring business operations problems that counteract strategic business goals (e.g. compliance with legal and regulatory provisions, high service level) or that cause high costs have not regularly been reported. Nevertheless, there have been critical voices on the quality of product data, particularly with regard to their cross-company use. For example, some distribution centers have been complaining about the accuracy of weights of newly launched products (i.e. of related logistic units). Such defects of logistics data can result in additional costs due to repackaging and restocking if tolerance range for pallet weights is exceeded.

Noticing these complaints, together with the growing awareness that product data quality is of particular importance at the interfaces with external partners (e.g. customer companies, data pool providers, logistics service providers), Beiersdorf's central Data Process Management unit initiated a project aiming at a) the identification of business critical data defects, and b) the specification of data quality metrics for monitoring these defects. The project plan comprised the following phases:

- Phase I: *Scoping*. Identification of data attributes and data quality dimensions deemed business critical.
- Phase II: *Interviews*. Identification of business operations problems caused by defects in the attributes focused (interviews with representatives from various actors of Beiersdorf's ecosystem).
- Phase III: *Analysis*. Consolidation of interview results and identification of critical data defects.

- Phase IV: *Specification*. Specification of data quality metrics for monitoring critical data defects.

Identification of critical data defects

As the product data model of Beiersdorf's PLM system comprises over 800 attributes, in Phase I the project team grouped these attributes in data clusters and made a pre-selection of 20 clusters (comprising about 100 attributes) to be examined in the interviews. Then four data quality dimensions were selected, namely *Accuracy* (i.e. data values correct?), *Completeness* (i.e. data values existent?), *Consistency* (i.e. same data values for an attribute processed by different systems?), and *Timeliness* (i.e. data values always timely available?). The team selected data clusters (e.g. bill of material, GTIN) and data defect types (e.g. ambiguous, incomplete, out-dated, wrong value) that were mentioned in complaints about product data quality. For example, GTIN inconsistencies could occur due to data changes in regional ERP systems (see Fig. 3). The dimensions were only used to structure the interviews, i.e. the team prepared several questions for each dimension that have been discussed for each selected attribute (e.g. 'Are there any wrong or inaccurate data?', 'Are there any out-dated data?', 'Are there any inconsistent data regarding data in the PLM system?').

In Phase II, interview partners were selected and interviews were conducted (see Table 1). Thereby, important parts of the supply chain as well as the complexity of the product portfolio (by taking into account products for the mass market, complex products, and both self-produced and externally produced products) could be covered well by seven interviews. Moreover, all interviewees have been with Beiersdorf for years (some of them for more than 20 years) and are well versed in the company's data, processes and systems. Interview participants were invited by the head of Supply Chain Technology (see Fig. 1) in order to highlight the interviews' significance. All interviewees supported the collaborative data defect identification by providing and discussing issues from business processes they are involved in. Complaints regarding other actors of the ecosystem could not be observed. Even for issues where required data is not provided timely in some cases (see [10] in Fig. 2), interviewees of the impaired party (production plant in the example) strived for the identification of mutual process improvements.

The interviews were based on an interview guide containing the grouped attributes for each cluster plus questions for each data quality dimension. Other questions referred to the interviewees' assessment of the current data quality in general, their confidence in the correctness of data, and other issues or problems not covered by the interview guide (e.g. further attributes). Each business operations problem identified was described, the data attribute and the defect type causing the problem were documented, and the frequency and business impact (i.e. costs, service level impairment, risk of complaints) of the problems were assessed using a scale ranging from 1 to 5. Given examples on how to interpret the respective rating scales ensured an approximately equidistant interpretation of the scale levels and a comparable use of the scales in different interviews. Table 2 summarizes the assessment results of all interviews and shows data attributes and corresponding issues that were mentioned during the interviews. For reasons of confidentiality Table 2 does not distinguish between particular interviews and issues and only provides means of frequency and impact ratings.

In Phase III the interview findings were consolidated in order to be able to identify those data clusters that are critical from the perspective of the entire ecosystem (see Table 2). As a result, the following seven issues could be identified.

- *Missing temperature conditions for transportation*. There were cases where, due to extreme outdoor temperatures, products had been delivered frozen, leading to customer complaints and costs for return, quality inspection, and new delivery. The configuration of the PLM system was lacking product specific information on ambient temperatures recommended for transport.
- *Wrong format of product marking (expiry date)*. Like many other companies, Beiersdorf has had in some cases problems using the correct format for indicating the expiry date on product packages. These formats differ from country to country. As there is no complete central documentation containing all valid formats for being looked up, provisions on the correct format must be often researched.
- *Missing dangerous goods indicator*. Beiersdorf has always made sure that dangerous goods (e.g. deodorant sprays) are marked and labeled accordingly, and so missing or wrong dangerous goods indicators have not been reported so far. However, products for special marketing campaigns (e.g. displays offering a bag with a shampoo and a deodorant spray) mostly are not packed in the plants where their components are produced but in the distribution centers. The need for dangerous goods indication then has to be identified by researching through the combined products' bills of materials. In several interviews this manual process was described as laborious and bearing risks, even if no concrete cases of failure were reported.
- *Missing or wrong GTIN*. There were cases where GTINs for logistic units (e.g. shrink, packaging unit, pallet) were missing, wrong, not unique (for example,

Table 2 Assessment results of interviews

Data attribute	Data cluster	Issues	Frequency	Impact
Gross weight	Logistics data	8	2.47	0.99
Country artworks	Artworks, technical drawings	7	2.60	0.53
Dangerous goods indicator	Dangerous goods indicator	5	2.33	0.97
GTIN (packaging unit)	GTIN	5	1.40	1.80
Measurement	Logistics data	5	2.42	0.55
Status	Bill of material	5	3.25	0.44
Temperature conditions	Temperature conditions	5	1.75	2.25
GTIN (piece level)	GTIN	4	1.25	1.00
GTIN (shrink level)	GTIN	3	2.00	1.55
Packing data	Bill of material	3	4.67	0.67
Product formula	Bill of material	3	2.67	0.33
Technical drawings	Artworks, technical drawings	3	4.50	0.67
Format of product marking	Format of product marking	2	4.50	1.67
Gross weight (pallet level)	Logistics data	2	3.00	0.33
Material cluster	Material data (basic data)	2	1.00	0.84
Material description	Material data (basic data)	2	2.00	0.50
Material number	Material data (basic data)	2	3.00	0.33
Stacking plan	Logistics data	2	1.50	2.33
Maximum storage period	Storage Data	1	3.00	0.55
Sub process cluster	Material data (basic data)	1	1.00	0.67

product GTINs used also for shrinks), or not consistent with the data transmitted to the GS1 data pool, potentially leading to problems in product distribution and lowering the service level.

- *Inaccurate or not timely available logistics data*. Calculated product weight (for example, volume multiplied by density) has not always been replaced in the PLM system by actual values determined after production according to final adaptations. In case of products with a transparent package (for example, shampoo filled in a clear bottle) sometimes a little bit more content than indicated on the label is filled in up to the bottles cap (otherwise the bottle looks somewhat empty). Although such changes in content usually is within the tolerance range of 20% as demanded by GS1, logistics processes might be affected by pallet weights being too high. Besides, information on product dimensions requested by wholesalers and retailers has not always been available in time.
- *Bill of material not timely released*. There were cases where bills of material had not been available in time or had been changed several times after the actual release date, leading to delays in the production process, increased effort in production plants, and potentially lower service level.
- *Documents not timely available*. There were cases where artworks and technical drawings had not been available in time, leading also to delays in the production process, increased effort in production plants, and potentially lower service level.

The identification of these seven problems constituted the starting point of the specification of data quality metrics, with two modifications being made. First, the problem of missing temperature conditions for transportation was not considered to be a data quality issue, as there was simply not enough information available for defining suitable thresholds. Second, in the course of an interim presentation, customs relevant data could be identified as another critical cluster. The data quality metrics were then extended by adding validation rules on timeliness for the attributes *commodity code* and *country of origin*. Table 3 lists the data defects (i.e. affected data attribute(s) and data quality dimension) deemed critical and assigns them to the data flows within Beiersdorf's ecosystem as shown in Fig. 2. The dimension *Change Frequency* was added in the analysis phase, as several issues referred to attributes (mostly entries and status values of bills of material) that were modified too often after the data's actual release.

Specification of data quality metrics

The aim of Phase IV was to specify data quality metrics allowing to monitor the data defects identified in the course of the interviews. Measured values of data quality metrics were normalized to the interval supposed to be interpreted as follows:

Table 3 Critical data defects and business impact

Data flow	Data cluster/ attribute	Data quality dimension	Business impact
[2]	Bill of material	Timeliness, Change Frequency	• Additional costs due to new production, if modifications of the bill of material make existing products unusable. • Service level impaired due to delays in production.
	Format of product marking	Accuracy, Completeness, Change Frequency	• Additional costs due to extra work (e.g. gathering information, communication). • Additional costs due to rework, if defect is not identified before start of production. • Risk of being fined by regulatory authorities if defect is not detected before shipment of goods.
[3]	GTIN	Consistency	• Service level impaired if inconsistencies are not detected before shipment of goods. • Risk of being fined by regulatory authorities if defect is not detected before shipment of goods.
	Logistics data	Accuracy	• Service level impaired if inconsistencies are not detected until shipment of goods. • Risk of being fined by regulatory authorities if defect is not detected before shipment of goods.
[4]	GTIN	Accuracy, Consistency	• Additional costs due to need for new artwork if defect is not detected before production or distribution.
[6]	Dangerous goods indicator	Completeness	• Risk of being fined by regulatory authorities if defect is not detected before shipment of goods.
	Logistics data	Timeliness	• Costs due to lost revenues.
[7]	Bill of material	Timeliness, Change Frequency	• Additional costs due to new production or rework, if modifications of the bill of material make existing products unusable. • Service level impaired due to delays in production.
[8]	Dangerous goods indicator	Completeness	• Additional costs due to extra work (e.g. gathering information, communication).
	GTIN	Accuracy, Completeness, Change Frequency	• Additional costs due to extra work (e.g. gathering information, communication). • Risk of being fined by regulatory authorities if defect is not detected before shipment of goods.
	Logistics data	Accuracy, Completeness, Change Frequency	• Additional costs due to extra work (e.g. gathering information, communication), particularly regarding GTINs for logistic units.
[10]	Artworks, technical drawings	Timeliness	• Service level impaired due to delays in production.
[11]	Dangerous goods indicator	Completeness	• Risk of being fined by regulatory authorities if defect is not detected by logistics service provider.
[12]	GTIN	Accuracy, Consistency, Completeness	• Additional costs due to extra work (e.g. gathering information, communication), particularly regarding GTINs for logistic units.
	Logistics data	Accuracy, Change Frequency	• Additional costs due to extra work (e.g. gathering information, communication), particularly regarding gross weight.
[13]	Dangerous goods indicator	Completeness	• Risk of being fined by regulatory authorities if defect is not detected by logistics service provider.
	Logistics data	Accuracy	• Additional costs due to repackaging and restocking if tolerance range is exceeded.
[14]	Customs relevant data	Timeliness	• Additional costs due to extra work (e.g. gathering information, communication). • Service level impaired due to delays in production.
[15]	Artworks, technical drawings	Timeliness	• Service level impaired due to delays in production.

- Best value (1): No data object validated contains a critical defect.
- Worst value (0): Each data object validated contains at least one critical defect.

The data clusters identified to be the cause of critical business problems constitute the structure of the metric system (i.e. one metric for each cluster). In order to be able to aggregate and compare measured values of metrics at a later stage, a calculation formula was determined that is applicable to all metrics (see below). For the same reason, a uniform structure for all validation rules was specified. A validation rule checks if a data object (representing the set of data attributes representing one product) shows all the characteristics defined by the rule (output value: 0) or if at least one criterion is not fulfilled (output value: 1).

$$r(d) = \begin{cases} 1, & \text{if the checked data object } d \text{ does not meet all criteria defined by } r \\ 0, & \text{if the checked data object } d \text{ meets all criteria defined by } r \end{cases}$$

The value of a data quality metric m is calculated by applying all $|R|_m|$ validation rules specified for m (R: set of all rules defined for a data class D) to those $|D|_{r_i}|$ data objects $d_j \in D$ which are defined to be checked by a specific rule r_i. For each validation rule r_i the weighting factor ω_i determines the rule's impact on the measured value of the metric m.

$$m(D) = 1 - \frac{\sum_{i=1}^{|R|_m|} \left(\frac{\sum_{j=1}^{|D|_{r_i}|} \omega_i r_i(d_j)}{|D|_{r_i}|} \right)}{\sum_{i=1}^{|R|_m|} \omega_i}, \text{with } \omega_i \geq 0$$

The formula's complexity results from the intention to be able to specify for each rule r_i as precisely as possible the subset $D|_{r_i}$ of D containing those $|D|_{r_i}|$ data objects in which defects can occur at all with respect to the data defect represented by r_i. Table 4 (see Appendix) shows the seven specified product data quality metrics and 32 validation rules. The rules are implemented in a reporting tool that periodically analyses all product data objects and calculates the metric values. The metrics are to be evaluated monthly.

Findings and discussion

Few complaints about product data quality in general and the result of conducted interviews indicate that Beiersdorf is dealing with a relatively small number of business critical product data defects which certainly is due to the good organization and technical support of the company's master data management. Inconsistent intra-company product data, duplicate data, or wrong positions in bills of material can largely be avoided by measures such as centralized assignment of unique identifiers, integration of product data maintenance processes already in the product innovation process (see MDW in Fig. 3), or clear specification of roles and responsibilities for product data maintenance. Another indicator for high maturity of Beiersdorf's master data management organization is the commitment of all interviewees regarding the data defect identification and metric specification process. All interviewees understood and supported the project's objectives, provided valuable input and offered additional support for answering further questions and for reviewing results.

Nevertheless, critical data defects do occur at Beiersdorf, particularly with regard to data exchanged with other businesses. Although the pre-selection of data clusters by the Data Process Management team (cf. Phase I of the project) leaded to a focus of product data that is used by multiple partners of the ecosystem, the interviewees did not raise any locally used attributes. Only customs relevant data was added which is used by multiple partners as well. Fortunately, many of the defects (e.g. missing dangerous goods indicator, wrong indication of pallet weights) are detected in the course of manual inspections done by experienced personnel. Well-defined cross-departmental or cross-divisional data management processes (e.g. maintenance of GTINs, design of artworks) have not been sufficiently established yet. Continuous monitoring of the data defects identified with the help of the data quality metrics specified will show if planned measures (e.g. new workflows, revised GTIN allocation process, redesigned process for artwork production) are able to sustainably improve product data quality.

To realize the specified data quality metrics system, Beiersdorf has been conducting a project for implementing specified metrics that comprises four phases: In a first phase, validation rules for measuring logistics data quality (see Table 4, VR24–VR30) and a reporting system have been implemented. The second phase comprises validation rules for measuring GTIN's and bill of material's change frequency, consistency and timeliness (VR01–VR03, VR20–VR 23, VR28–VR30). In contrast to the first phase, validation rules in Phase 2 require access to time related metadata (i.e. change protocols) and to further systems (e.g. VR19). In Phase 3, validation rules for monitoring GS1 compliance (e.g. VR VR09–VR18) will be implemented. And Phase 4 will provide remaining validation rules such as VR05–VR07 for monitoring timely availability of documents and customs data (VR04), for monitoring consistent dangerous goods indicators (VR08) and for product marking format (VR31, VR32).

Implications for practice and research

The case study at Beiersdorf describes the process of identifying business critical product data defects and proposes a set of seven data quality metrics that allow to monitor such defects. To measure these metrics, 32 validation rules are specified using only commonly used (i.e. not specific to Beiersdorf) product data attributes. They may be used also by other companies as elements of their own metrics systems. Thus, an obvious implication for practice is to use and evaluate the specified metrics as elements of their own data quality metrics systems and to implement the validation rules within their measuring and reporting tools.

As is the case with all case studies, whether the results (here: product data defects, data quality metrics) can be generalized needs to be found out by conducting more case studies with similar companies and by applying proven theories to get a better understanding of the interrelations identified in this case study. An important implication for research therefore is to motivate scientists to evaluate and adapt the metrics and validation rules specified for Beiersdorf, and to investigate organizational (e.g. degree of collaboration) and technical (e.g. complexity of data models) factors impacting the quality of shared product data.

Another implication for research is due to potential dependencies between the impact of data defects and the time of their detection regarding the data lifecycle. Wrong logistics data (e.g. a wrong pallet weight) may cause only little additional costs (for reweighing and repackaging) when the defect is detected at a distribution center.

When goods arrive at the customer's site, however, this data defect may bring about high additional costs (for return of the goods and lowered service level). This means that not just the type of data defect but also the time a data defect is detected seem to have an effect on its business impact. Modeling and monitoring a data lifecycle perspective of a company's ecosystem comprising all create, read, update, deactivate and archive activities would facilitate model-based analysis of data flows and data maintenance processes and help identify weak points in data maintenance processes (e.g. missing responsibilities, cycles, delays).

Acknowledgements This research was conducted within the Competence Center Corporate Data Quality (CC CDQ) at the Institute of Information Management at the University of St. Gallen. The CC CDQ is part of the research program Business Engineering at the University of St. Gallen (BE HSG).

Appendix

Table 4 Data quality metrics and validation rules

Metrics and validation rules	Explanation
Metric 'Bill of material' (BoM)	

Validation rule VR01
`if (CurrentTime() + threashold) > FirstDemand(d)`
`    if d.BoM.Status != 'released' : return 0`
`end if`
`return 1`
	The BoM (i.e. its positions) should only be modified until *threshold* days before the first demand in order to guarantee the availability of ingredients.

VR02
`for each position p in d.BoM.Positions`
`    if d.BoM.Status == 'ready'`
`        if NumberOfChanges(p) > 1 : return 0`
`    end if`
`end for`
`return 1`
	Modifications of BoM positions should be avoided in order to ensure performance of processes using the BoM (e.g. production in plants).

VR03
`if NumberOfChanges(d.BoM.Status, 'released') > 1 : return 0`
`return 1`
	Modifications of a BoM after its release should be avoided.

Metric 'Customs relevant data'	

VR04
`if (CurrentTime() + threashold) > FirstDemand(d)`
`    if d.CommodityCode   == BLANK : return 0`
`    if d.CountryOfOrigin == BLANK : return 0`
`end if`
`return 1`
	In order to avoid logistic delays, customs relevant data (i.e. the commodity code and the country of origin) should be present *threshold* days before the first demand.

Metric 'Documents'	

VR05
`if (CurrentTime() + threashold) > FirstDemand(d)`
`    if d.Artwork.Status   != 'released' : return 0`
	A product's artwork should be present (and released) *threshold*

Table 4 (continued)

Metrics and validation rules	Explanation

`    if d.Artwork.Document == BLANK      : return 0`
`end if`
`return 1`
	days before the first demand.

VR06
`if d.PackagingFamily.Status == 'released'`
`    if d.TechnicalDrawing.Status != 'released' : return 0`
`end if`
`return 1`
	For each specification of a product's packaging material a technical drawing should be available.

VR07
`plm = LastChange(PLM.d.BoM.Status, 'released')`
`mdw = LastChange(MDW.d.BoM.Status, 'released')`
`if plm > mdw + threshold : return 0`
`return 1`
	If a BoM is released in the MDW system it should be released in the PLM system as well (within *threshold* days).

Metric 'Dangerous-goods indicators' (DGI)	

VR08
`if d.BoM.Status == 'released'`
`    if d.DGI == 'no'`
`        for each position p in d.BoM.Positions`
`            if p.d.DGI == 'yes' : return 0`
`        end for`
`    end if`
`end if`
`return 1`
	If a product contains dangerous goods, its DGI flag must be set.

Metric 'GTIN'	

VR09 – VR12 (check for GS1 compliance for products (see below) and for logistic units shrink, packaging unit and pallet)
`for each data object d2`
`    if d.GTIN == d2.GTIN`
`        if GS1(d, d2) == FALSE : return 0`
`    end if`
`end for`
`return 1`
	GS1 defines a gross weight tolerance for products (and logistic units) with the same GTIN of 20%.

VR13
`for each data object d2`
`  if d.GTIN == d2.GTIN`
`    if d.Shrink.Item        != d2.Shrink.Item        : return 0`
`    if d.PackagingUnit.Item != d2.PackagingUnit.Item : return 0`
`    if d.Pallet.Item        != d2.Pallet.Item        : return 0`
`  end if`
`end for`
`return 1`
	Two products with the same GTIN are not allowed to have a different number of items on secondary logistic levels.

VR14
`if d.GTIN != ArtworkGTIN(d) : return 0`
`return 1`
	The data values of a product's GTIN stored in the PLM system and printed on the product's artwork should be equal.

VR15 – VR17 (for product level (see below), and logistic units shrink and packaging unit)
`if d.GTIN == BLANK : return 0`
`return 1`
	GTINs should be defined for all products.

VR18
`for each data object d2`
`    if d.GTIN               == d2.Shrink.GTIN        : return 0`
`    if d.GTIN               == d2.VE.GTIN            : return 0`
`    if d.GTIN               == d2.PackagingUnit.GTIN : return 0`
`    if d.Shrink.GTIN        == d2.PackagingUnit.GTIN : return 0`
`    if d.Shrink.GTIN        == d2.Pallet.GTIN        : return 0`
`    if d.PackagingUnit.GTIN == d2.Pallet.GTIN        : return 0`
`end for`
`return 1`
	GTINs of all products and logistic levels must be unique. The PLM system already guarantees unique GTINs on product level.

VR19
`for each regionalSAP sap`
`    if d.GTIN    != sap.d.GTIN               : return 0`
`    if d.SH.GTIN != sap.d.Shrink.GTIN        : return 0`
`    if d.VE.GTIN != sap.d.PackagingUnit.GTIN : return 0`
`    if d.P.GTIN  != sap.d.Pallet.GTIN        : return 0`
`end for`
`return 1`
	GTIN information is transferred to GS1 not from the central PLM system but from regional SAP systems. Therefore, GTINs from PLM should be consistent with GTINs from regional SAP

Table 4 (continued)

Metrics and validation rules	Explanation
	systems.

VR20 – VR23 (for product level (see below), and logistic units shrink, packaging unit and pallet)
`if NumberOfChanges(d.GTIN) > 1: return 0`
`return 1`
	A GTIN change causes effort in various processes (e.g. artwork production) and should be avoided.

Metric 'Logistics data'	

VR24 – VR27 (for product level (see below), and logistic units packaging unit and pallet)
`if d.GrossWeight == BLANK : return 0`
`if d.Height == BLANK : return 0`
`if d.Lenght == BLANK : return 0`
`if d.Width == BLANK : return 0`
`if d.Volume == BLANK : return 0`
`return 1`
	Logistic data should be defined for all products.

VR28 – VR30 (for logistic units shrink (see below), packaging unit and pallet)
`if NumberOfChanges(d.Shrink.Items) > 1 : return 0`
`return 1`
	Pieces per shrink (and further logistic levels) should not be changed after release.

Metric 'Product marking format' (PMF)	

VR31
`if d.PMF == 'no requirements' : return 0`
`else : return 1`
	Missing formats of product marking might cause legal complaints.

VR32
`if NumberOfChanges(d.PMF) > 1 : return 0`
`return 1`
	The format of product marking should be defined only once as corresponding legal regulations rarely change.

The validation rules listed in Table 4 partially contain functions that provide metadata for data fields or that execute special validation procedures:

- *CurrentTime*. Provides the current time.
- *FirstDemand*. Provides the timestamp a series of products is produced for the first time.
- *NumberOfChanges*. Provides the number of changes of a value of a data attribute.
- *GS1*. Checks if the difference in the gross weights of two products is more than 20% (if this is the case, the products need to be assigned with different GTINs).
- *ArtworkGTIN*. Provides the GTIN indicated on a product's artwork (application of barcode image processing algorithms).

References

Baskerville, R. L. (1997). Distinguishing action research from participative case studies. *Journal of Systems and Information Technology, 1*(1), 24–43.

Batini, C., Barone, D., Mastrella, M., Maurino, A., & Ruffini, C. (2007). *A Framework and a Methodology for Data Quality Assessment and Monitoring.* Paper presented at the 12th International Conference on Information Quality, Cambridge.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), 1–52.

Batini, C., & Scannapieco, M. (2006). *Data quality, concepts, methodologies and techniques*. Berlin: Springer.

Beck, R., & Weitzel, T. (2005). Some economics of vertical standards: integrating SMEs in EDI supply chains. *Electronic Markets, 15* (4), 313–322.

Beiersdorf. (2010). *Annual Report 2010*. Hamburg: Beiersdorf.

Caballero, I., Calero, C., Piattini, M., & Verbo, E. (2008). *MMPro: A methodology based on ISO/IEC 15939 to draw up data quality measurement processes.* Paper presented at the 13th International Conference on Information Quality, Cambridge.

Cachon, G. P., & Lairiviere, M. A. (2000). Supply chain inventory management and the value of shared information. *Management Science, 46*(8), 1032–1048.

Cai, Y., & Ziad, M. (2003). *Evaluating Completeness of an Information Product.* Paper presented at the 9th Americas Conference on Information Systems, Tampa.

Chen, F., Drezner, Z., Ryan, J. K., & Simchi-Levi, D. (2000). Quantifying the bullwhip effect in a simple supply chain: the impact of forecasting, lead times, and information. *Management Science, 46*(3), 436–443.

English, L. P. (1999). *Improving data warehouse and business information quality.* New York: Wiley.

Eppler, M. J., & Helfert, M. (2004). *A Classification and Analysis of Data Quality Costs.* Paper presented at the 9th International Conference on Information Quality, Cambridge.

Fan, W., Geerts, F., Jia, X., & Kementsietsidis, A. (2008). Conditional functional dependencies for capturing data inconsistencies. *ACM Transactions on Database Systems, 33*(2), 1–48.

Gelinas, U. J., & Dull, R. B. (2009). *Accounting information systems* (8th ed.). Mason: South-Western.

Goh, C. H., Bressan, S., Madnick, S. E., & Siegel, M. (1999). Context interchange: new features and formalisms for the intelligent integration of information. *ACM Transactions on Information Systems, 17*(3), 270–293.

Gorla, N., & Krishnan, K. (2002). *Information quality: Modeling accuracy- timeliness in transaction processing systems.* Paper presented at the 8th Americas Conference on Information Systems, Dallas.

GS1 (2010). *GS1 general specifications, Version 10*: Global Standards One.

Heinrich, B., & Klier, M. (2009). *A novel data quality metric for timeliness considering supplemental data.* Paper presented at the 17th European Conference on Information Systems, Verona.

Heinrich, B., Klier, M., & Kaiser, M. (2007). *DQ metrics: A novel approach to quantify timeliness and its application In CRM.* Paper presented at the 12th International Conference on Information Quality, Cambridge.

Heinrich, B., Klier, M., & Kaiser, M. (2009). A procedure to develop metrics for currency and its application in CRM. *Journal of Data and Information Quality, 1*(1), 1–28.

Hipp, J., Müller, M., Hohendorff, J., & Naumann, F. (2007). *Rule-based measurement of data quality in nominal data.* Paper presented at the 12th International Conference on Information Quality, Cambridge.

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM, 53*(01), 148–152.

Kulp, S. C., Lee, H. L., & Ofek, E. (2004). Manufacturer benefits from information integration with retail customers. *Management Science, 50*(4), 431–444.

Lau, J. S. K., Hunag, G. Q., & Mak, K. L. (2002). Web-based simulation portal for investigating impacts of sharing production information on supply chain dynamics from the perspective of inventory allocation. *Integrated Manufacturing Systems, 13*(5), 345–358.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: the Bullwhip effect. *Management Science, 43*(4), 546–558.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management, 40*(2), 133–146.

Legner, C., & Schemm, J. (2008). Toward the inter-organizational product information supply chain—evidence from the retail and consumer goods industries. *Journal of the Association for Information Systems, 9*(4), 119–150.

Lehmann, H. (2003). An object-oriented architecture model for international information systems? *Journal of Global Information Management, 11*(3).

Loshin, D. (2008). *Master data management*. Burlington: Kaufmann.

Madnick, S. E. (1995). Integrating information from global systems: dealing with the "On- and Off-Ramps" of the information superhighway. *Journal of Organizational Computing and Electronic Commerce, 5*(2), 69–82.

Nakatani, K., Chuang, T.-T., & Zhou, D. (2006). Data synchronization technology: standards, business values and implications. *Communications of AIS, 17*(1), 962–994.

Nicolaou, A. I., & McKnight, D. H. (2006). Perceived information quality in data exchanges: effects on risk, trust, and intention to use. *Information Systems Research, 17*(4), 332–351.

Otto, B., Hüner, K. M., & Österle, H. (2009). *Identification of business oriented data quality metrics.* Paper presented at the 14th International Conference on Information Quality, Potsdam.

Price, R., Neiger, D., & Shanks, G. (2008). Developing a measurement instrument for subjective aspects of information quality. *Communications of AIS, 2008*(22), 49–74.

Price, R., & Shanks, G. (2005). A semiotic information quality framework: development and comparative analysis. *Journal of Information Technology, 20*(2), 88–102.

Redman, T. C. (1996). *Data quality for the information age*. Boston: Artech House.

Schemm, J. W. (2008). *Zwischenbetriebliches Stammdatenmanagement: Lösungen für die Datensynchronisation zwischen Handel und Konsumgüterindustrie.* PhD thesis, University of St. Gallen, Institute of Information Management, St. Gallen, Switzerland.

Scholz, R. W., & Tietje, O. (2002). *Embedded case study methods. Integrating quantitative and qualitative knowledge*. Thousand Oaks: Sage.

Shankaranarayanan, G., & Cai, Y. (2006). Supporting data quality management in decision-making. *Decision Support Systems, 42* (1), 302–317.

Sheu, C., Yen, H. R., & Chae, B. (2006). Determinants of supplier-retailer collaboration: evidence from an international study. *International Journal of Operations & Production Management, 26*(1), 24–49.

Smith, H. A., & McKeen, J. D. (2008). Developments in practice XXX: master data management: salvation or snake oil? *Communications of the AIS, 23*(4), 63–72.

van der Blonk, H. (2003). Writing case studies in information systems research. *Journal of Information Technology, 18*(1), 45–52.

Vermeer, B. H. P. J. (2000). *How important is data quality for evaluating the impact of EDI on global supply chains?* Paper presented at the 33rd Annual Hawaii International Conference on System Sciences, Maui.

Walsham, G. (1995). Interpretive case studies in IS research: nature and method. *European Journal of Information Systems, 4*(2), 74–81.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86–95.

Wang, R. Y., Lee, Y. W., Pipino, L. L., & Strong, D. M. (1998). Manage your information as a product. *Sloan Management Review, 39*(4), 95–105.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.

Wang, S., Archer, N. P., & Zheng, W. (2006). An exploratory study of electronic marketplace adoption: a multiple perspective view. *Electronic Markets, 16*(4), 337–348.

Weber, K., Otto, B., & Österle, H. (2009). One size does not fit all—a contingency approach to data governance. *ACM Journal of Data and Information Quality, 1*(1), 4:1–4:27.

Yin, R. K. (2002). *Case study research: design and methods* (3rd ed.). Thousand Oaks: Sage.

Zhao, H. (2007). Semantic matching across heterogeneous data sources. *Communications of the ACM, 50*(1), 44–50.