A Reference Model for Product Data Profiling in Retail ERP Systems

Rolf Krieger and Christian Schorr

Institute for Software Systems, University of Applied Sciences Trier, Birkenfeld, Germany

Keywords: Data Quality, Data Profiling, Exploratory Data Analysis, Product Data, Enterprise Resource Planning.

Abstract: Due to the high volume of data and the increasing automation in retail, more and more companies are dealing with procedures to improve the quality of product data. A promising approach is the use of machine learning methods that support the user in master data management. The development of such procedures demands error-free training data. This means that product data must be cleaned and labelled which requires extensive data profiling. For typical retail company data bases with usually complex and convoluted structures this exploration step can take a huge and expensive amount of time. In order to speed up this process we present a reference model and best practices for the systematic and efficient profiling and exploration of product data.

1 INTRODUCTION

Product data quality is an increasingly important topic in the retail industry. The automation of purchasing, sales, delivery and inventory processes can only succeed if the referenced product data are error-free. Especially for e-commerce where the decision to buy a product can only rely on its virtual presentation in a web shop, correct product descriptions are of utmost importance. Moreover, highly automated storage centres where robots move products from the shelves to the packaging area require precise data about physical dimensions and weight in order to ensure a flawless delivery to the waiting customer. If the quality of the product data is poor, delays in delivery, losses due to incorrect prices, additional work etc. are the result.

Manual data quality control is nearly impossible due to the vast amount of products. Current semi-automated quality control systems based on data quality rules are prone to miss erroneous data due to the vast amount of attributes, tables and their complex relations. Machine learning provides a promising approach to support the user to ensure data quality. But major prerequisites for machine learning are high quality test and training data (Domingos, 2012). Consequently, at the beginning of a machine learning project the quality of product data has to be estimated and product data have to be cleaned. In addition domain understanding is of importance and can only be gained by a careful and detailed exploration of the data. Intensive data profiling is therefore necessary in order to gather meta data and to get acquainted with the product data.

To ease this central task of product data quality management, we propose a reference model which defines a reference process and a high-level architecture model of a data profiling tool taking into account the specific properties and requirements of product data. The reference process should allow to identify critical data records efficiently. Based on our results the process of profiling product data should be accelerated significantly.

The paper is structured as follows: Section 2 describes the goals, tasks and state of the art of data profiling while section 3 explains product data. In Section 4, we introduce our reference process for profiling product data. In section 5, we present best practices to support the profiling steps and illustrate them using examples. Finally, we summarize the main results and give an outlook on our further research.

2 STATE OF THE ART

Data profiling is the application of data analysis with the purpose of determining information about the content, structure and quality of data. According to (Olson, 2008), the input of the profiling process are data and meta data about them. Meta data describe the structure, content, history and quality of the data. In practice meta data are usually incomplete. The output of the data profiling process are information about erroneous data and complete meta data. Thus data profiling is carried out to gain business understanding and

Krieger, R. and Schorr, C.
A Reference Model for Product Data Profiling in Retail ERP Systems.
DOI: 10.5220/0007953303170324
In *Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019)*, pages 317-324
ISBN: 978-989-758-377-3

to provide information for data cleaning. Analysing the structure of the data involves checking their consistency and format. Computing descriptive statistical information like minima, maxima and percentages as well as determining data types and lengths falls under this category. Profiling the data content identifies specific properties concerning missing values or other errors. A crucial part of data profiling especially with regard to subsequent feature engineering tasks in machine learning projects is to discover how different parts of the data are related to each other. Identifying embedded value dependencies and functional dependencies between attributes or tables as well as potential (foreign) keys are part of this data profiling aspect. An excellent overview about the topic can be found in (Abedjan et al., 2015). Data profiling is also related to data exploration. A survey on data exploration techniques can be found in (Guido et al., 2015; Di Blas et al., 2014). Both commercial and open-source software tools are available for data profiling and cleaning. Often vendors of commercial ERP systems also have a corresponding data quality tools in their portfolios. SAP sells its Information Steward, Oracle its Enterprise Data Quality and IBM offers InfoSphere Information Server. The company Informatica specialized in business analytics sells Informatica Data Profiling while SAS offers its DataFlux Management Studio. Apart from their comprehensive scope, the major advantage of these tools is their seamless integration into the respective ERP systems of their companies, which avoids migration and transformation problems. Their disadvantage is that they are not optimized for the profiling of product data in general. Beside these commercially available products a plethora of open-source tools exists. A few mature software packages are Metonome (Papenbrock et al., 2015), developed and managed by the University of Potsdam, Germany, Profiler (Kandel et al., 2012) and Talend Open Studio. The drawbacks of free data profiling software usually lie in their narrow focus on specific applications especially in case of more scientific origins. In contrast to these tools we propose a data profiling tool and reference process specifically taking into account domain knowledge about product data.

3 PRODUCT DATA

3.1 General Description

Product data are of utmost importance in retail companies. For technical and business reasons, products are classified into categories arranged in a multi-level merchandise category hierarchy. Moreover, there are a variety of different product types like simple product items, grouped products (e.g. displays, sets etc.) and generic products and their variants. Product attributes are referenced by many processes in purchasing, material requirements planning, inventory management and sales. Table 1 shows some attributes of product data in a standard ERP software. We can distinguish between descriptive and process-related attributes.

The values of descriptive attributes are the same for all locations (e.g. stores, warehouses) of a company whereas process attributes may be maintained differently. For example, the safety stock may be 10 pieces in one store and 15 pieces in another. Process-related attributes control the processes in a company and are decisive for process automation. Usually the product attributes provided in a standard ERP software do not meet all requirements of a company. Therefore, they are supplemented by custom-specific attributes, which can be either descriptive or process-related.

Mostly, product attributes are also classified in the ERP software thematically. Typical classes are basic data, logistic data, purchasing data etc. Basic data include the product description, packaging units, dimensions, volumes, gross and net weights, hazardous substance codes etc. Logistic data include safety stock, service level, delivery time, goods receipt processing time etc. Purchasing and sales data include prices, minimum order quantities, delivery periods etc.

The attributes that have to be maintained for a specific product item depend on its category and type. Consequently, the product categories and types have to be taken into account in the data profiling process.

Table 1: High-level classification of product attributes in a ERP system.

	descriptive	process-related
standard	- product description - category - volume - unit of measure ...	- reorder point - safety stock - goods receipt processing time ...
custom-specific	- alcoholic strength - description of ingredients ...	- age limit for sales - flags for logistic processes ...

3.2 Case Study: SAP Retail

To get an idea of the amount and complexity of product data managed in an ERP system, we conducted a case study in two European full-range retail companies. Both companies use SAP Retail as ERP system to support their business processes. Table 2 gives some information about the product items of the companies.

Table 2: Rounded number of product categories, product items and types of two retail companies.

	Company 1	Company 2
Main categories	12	90
Sub categories	200	3.200
Items	750.000	≥ 1.400.000
Product types	7	9

As in many other ERP systems, in SAP Retail the attributes of products are divided into data segments, e.g. basic data, purchasing data, sales data and logistic data and are managed in more than 10 tables. Table 3 shows the number of attributes of the basic data table. For instance, the basic data table of company 1 has 290 attributes. 80 attributes are in use. 42% of them are custom-specific.

Table 3: Number of attributes of the basic data table in SAP ERP of two companies.

Attributes	Company 1	Company 2
Total	290	245
standard	206	209
custom	84	36
standard (used)	46	43
custom (used)	34	20

3.3 Technical Challenges

Due to the results of our case study, profiling and cleaning of product data are difficult tasks. We identified the following challenges:

3.3.1 High Number of Product Items

Big retail companies have more than 1 million items in their ERP systems. Some of the attributes, e.g. logistic attributes, have to be maintained for each location. In case of a company operating 50 stores 50 million logistic data records have to be maintained.

3.3.2 High Number of Product Categories

Some larger retail companies have more than 3.000 product categories. The category to which a product item belongs often defines the attributes that have to be maintained. Mandatory fields in particular are defined by the product categories. The large number of product categories makes it difficult to identify erroneous product data records.

3.3.3 Different Product Types

There are a variety of different product types. Similar to the product categories they define the attributes that have to be maintained for each item. Consequently, the product items have to be partitioned by product types for data profiling.

3.3.4 Age of the Data

ERP systems are often in use for years. Due to the complex IT system landscape in a retail company, product data are archived rarely. Hence, a lot of outdated items are still kept in the ERP system.

3.3.5 Business Processes Optimization

For business, technical and organizational reasons, business processes are continuously optimized or adapted in retail. Changing a business process often requires the adjustment of process related attributes of the product data. Mostly, outdated items are not updated accordingly and do not fulfil current business and technical requirements.

3.3.6 Complex Data Model

Product data are managed in many data tables in a ERP system. In our case studies we identified more than 10 tables. Some of these tables have a lot of attributes (e.g. 80) of different types (e.g. numerical, character, dates etc.). In case of some attributes specific knowledge of application experts is required to understand their semantic and usage.

3.3.7 Changes to the Data Model

ERP systems are usually standard software. If a new version of the software is developed by the software manufacturer, this is often associated with changes to the data model. On the one hand attributes may be added, on the other hand existing attributes may be no longer used. In addition, the data model is adapted in a similar way by the customer. The data profiling process has to take all these changes into account.

4 REFERENCE MODEL

In order to enable an efficient profiling of product data despite the challenges and problems formulated in the last section, we have developed a reference model consisting of a reference process and an architectural model of a data profiling tool.

4.1 Reference Process

4.1.1 Data Partitioning

The main problem with profiling product data is the amount and variety of data. In order to achieve meaningful results efficiently, the data must be partitioned appropriately for data profiling. Table 4 shows the most important partitioning criteria by data segments.

Table 4: Partitioning criteria by product data segments.

Partitioning criteria	Product data segments		
1. Product type 2. Product category	Basic data		
3. Location, e.g. store, warehouse	Logistic data	...	
4. Business organisation	Sales data	Purchasing data	...

In general, product data can be partitioned by type and category. Especially, basic data can be partitioned by product type and category. In the case of logistic data, locations can be used for partitioning in addition. Purchasing and sales data can also be partitioned by business department. As is well-known, partitioning leads to smaller amounts of data, reduces data variability and simplifies the data profiling.

4.1.2 Data Profiling and Exploration

For each element of the partition, we suggest a hierarchical approach to data profiling as illustrated in Figure 1. Using this concept, the product data are analysed and evaluated at the merchandise category level in the first step. This allows to detect anomalies by comparing results for different categories with available meta data and expert knowledge. In addition, the results of different categories are compared with each other. Differences between categories often indicate anomalies caused by erroneous data.

Due to the large number of attributes and categories, heat maps are often used in this step to visualize the results. Examples are given in Section 5. The information obtained in this way is used to check, correct and extend existing meta data.

The data profiling process can thus be executed using a drill down approach. This means that the results are first analysed for main categories, then for categories, subcategories and finally for product items as illustrated in Figure 1. This allows the data profiling to be performed efficiently even with large amounts of data containing many attributes.

In addition to the hierarchical approach the reference process also proposes the order in which data segments should be considered. Figure 2 shows the order of data segments.

The basic data should be considered in the first step. Usually, they are independent of the locations, e.g. stores and warehouses. In addition, the assignment of the products to the stores and warehouses has to be considered. This is necessary because the assortments of the stores differ - not all products are sold in all stores. Afterwards, the knowledge gained is used to look at the logistic data of the products which depend on locations. In many cases, the products are transported to the stores via multi-level logistic networks. Therefore, the logistic data for the warehouses should be considered before the logistic data for the stores. Finally, the purchasing and sales data of products that are dependent on business partners are analysed.

For each step we propose to develop specific scripts to run through the reference process in a time and cost efficient way.

4.2 High Level Architecture of the Tool

To evaluate the reference process, we have developed a prototype of the tool as a package in the statistical programming language *R*. It offers functions and scripts for product data profiling of the different data segments and for the visualization of the results, especially for basic and logistics data. Figure 3 describes the high level architecture and the application of the package.

4.2.1 Data Import

Our tool contains scripts for importing the data and meta data, building the base for data exploration. In addition to the product data, we have also provided the option of loading change documents that describe changes of the product data depending on time and transaction. When loading, the data types of the ERP system are converted to the corresponding *R* data types. In order to automate this process, the data dictionary of the ERP system is accessed. The type mapping and the encoding of missing values are provided in an XML document. Basic data cleaning steps

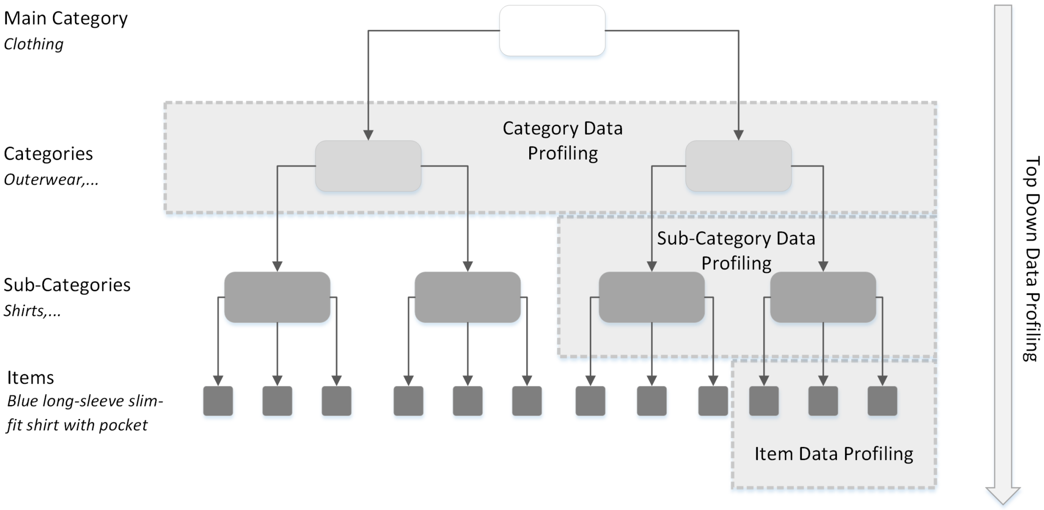

Figure 1: Description of the hierarchical top down data profiling process with example categories for a fashion retail data base.

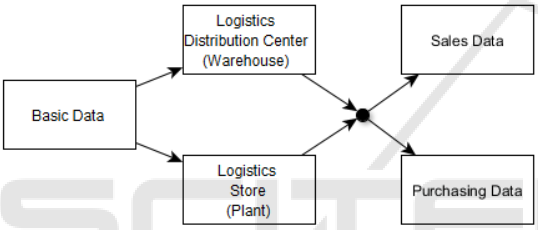

Figure 2: Order of data segments as considered in the data profiling reference process.

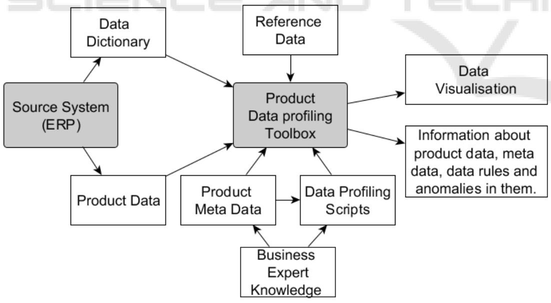
Data Dictionary
Reference Data
Source System (ERP)
Product Data profiling Toolbox
Data Visualisation
Information about product data, meta data, data rules and anomalies in them.
Product Data
Product Meta Data
Data Profiling Scripts
Business Expert Knowledge

Figure 3: High-level architecture model of our data profiling package based on *R*. Systems are shaded in grey.

are also performed. For instance, non-informative attributes (completely missing attributes and constant attributes) are identified and removed. The different encodings of missing values (e.g. 00.00.0000, 0.000, " " etc.) are replaced by the value NA.

4.2.2 Reference Data

Reference data are required for certain profiling steps. In particular the following reference data are required:

1. Merchandise Category Hierarchy.

   To support the hierarchical data profiling approach the merchandise category hierarchy of the company must be part of the reference data. Additionally, the standard classification systems UNSPSC (www.unspsc.org), eCl@ss (www.eclass.eu) and GPC (www.gs1.org) used in many companies should be provided.

2. Conversion Factors for Units of Measurement.

   Usually, the dimensions and volumes of products are given in different units of measurement. A conversion function allows the user to easily convert the values into a defined target unit using conversion factors provided by the reference data. Custom-specific types of packaging units like pallets containing different amounts of pieces are converted to the general basic unit pieces for example.

3. Dimensions and Weights of Standardized Packaging Units.

   Packaging units such as boxes and pallets that are referenced in product data are often standardised. If their dimensions and weights are known, detailed analyses can be carried out when profiling the product data. For instance, a comparison of the gross and net weight of a product in a particular packaging unit can be done in detail. Often this reveals erroneous data records.

4.2.3 Data Profiling Scripts

Our package provides functions for analysing single attributes, multiple attributes, and data rules. They form the basis for the scripts developed for each data segment. Bar charts and box plots are used to visualize the results of univariate analyses. For bivariate analyses scatter diagrams, heat maps, calendar-heat maps and bag plots are used. For each data segment a script defines the key figures and diagrams to support the profiling of product data. The package is written in *R* and uses custom-made code as well as standard libraries. It is tailored to be used with data in SAP for Retail format but can be adapted to work with other systems, too.

5 BEST PRACTICES

The scripts are created with the process knowledge of product data management experts and the technical knowledge of IT experts and correspond to best practices for the profiling of product data. In the following we describe some of the scripts and give some examples regarding the basic data segment.

5.1 Maintenance Status

An important step in data profiling is the analysis of the data segments that are created for each product item. Different segments are required depending on the product type and category. A heat map showing the frequency of data segments by type (or category) can help to understand the dependencies and to identify erroneous items. Typically there are some items with missing data segments.

5.2 Creation and Change Dates

As mentioned before, retail ERP systems often contain data about products that are obsolete and do not belong to the current product assortment. The creation date and the date of the last change of a data record provide important information about this. Our script determines the number of new product data records by day, week and month. Based on a multivariate analysis the merchandise categories are identified in which most products are created per day, week and month. In addition the frequencies of attribute values depending on the creation dates are analysed. This provides information about product items and categories that were created a long time ago and have not undergone any changes. Together with technical experts, it should be checked whether these products can be excluded from further data profiling. In many cases this leads to a considerable reduction in effort. In our case study it was determined that only about 20% of the products belong to the current assortment.

5.3 Merchandise Categories

The distribution of product categories and their subcategories can easily be visualized using a tree map as shown in Figure 4. It displays the grouping of products along their hierarchy levels. It is useful to get a quick graphical overview about the distribution of the product data by category. Additionally, it is helpful in determining whether products are categorized correctly.

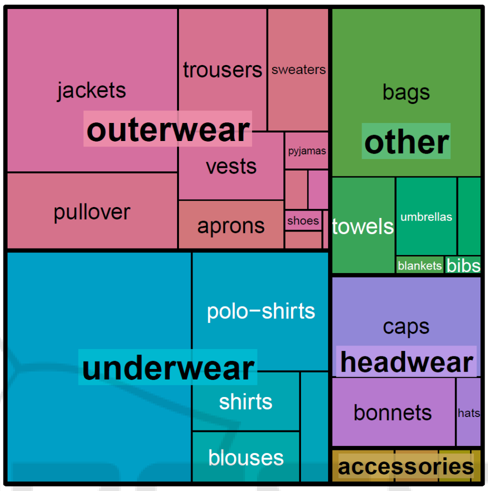

Figure 4: Tree map illustrating the frequencies of items by category and subcategory using textile product data. All articles are first grouped by category and then subdivided by subcategory. The category headwear for example has the subcategories caps, bonnets and hats.

The purpose of this analysis is to determine the main characteristics of the merchandise category hierarchy, the number of main categories, categories, and subcategories for example. As a rule, the individual categories differ in the number of subcategories they contain. The frequency of the subcategories per category can be visualized by a box plot. Outliers can also be detected and further investigated. Categories with a high number of subcategories and products must be further subdivided for data profiling.

5.4 Frequencies Product Items

The number of different product types and product categories as well as the frequencies of the product items per type and category are of importance. They allow us to derive the number and size of the partitions, which are needed to plan and execute the subsequent profiling steps.

5.5 Dimensions, Volumes and Weights

The dimensions, volume and weight of products are important for space planning in the shop and for the transportation planning of the products. These data are analysed taking into account the different merchandise categories and product types. Outliers are identified. By comparing the different merchandise categories, further anomalies are found. It should be noted that several packing units are often assigned to one product as shown in Figure 5. Product items having an infrequent combination of units are suspicious and should be analysed in detail.

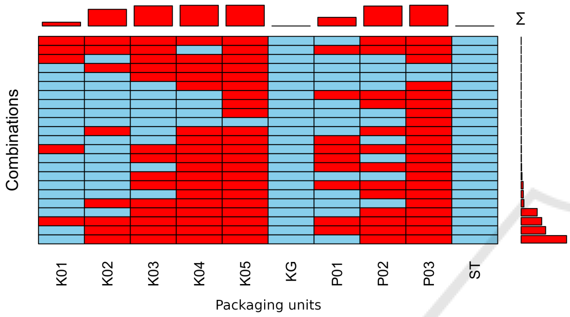

Figure 5: Frequencies of items by combinations of packing units. Red fields indicate that a packaging unit is used, while blue fields stand for unused packaging units. The right-hand bars depict the total amount of a combination for all items. The bars above show the total amount of each packaging unit for all items. The combinations in the upper rows with very small frequencies are possibly errors.

In addition, we have implemented various data rules (e.g. product dimensions are very small, very large, etc.) to identify erroneous product data records.

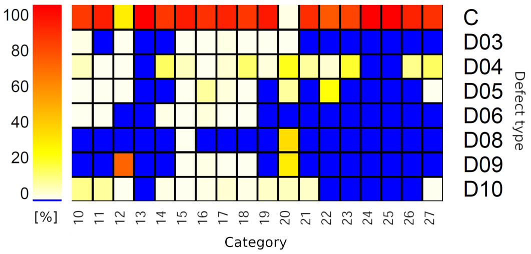

Figure 6: Heatmap of the percentage distribution of items by defect classes for each category. For category 12, about 20% of all items are correct, 60% display defect D09, while the remaining 20% belong to the defect classes D03, D04, D05 and D10, for example.

Figures 6 and 7 show examples of the results of such analyses. Figure 7 is the result of a drill-down by disaggregating the merchandise category.

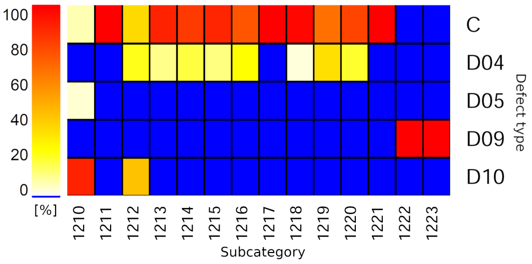

Figure 7: Heatmap of the percentage distribution of items by defect classes for each subcategory. The items of category 12 as seen in the previous figure are split into subcategories and analysed regarding the same defect classes. The detailed analysis shows that the subcategories 1210, 1212, 1222 and 1223 contain the majority of all errors in category 12. All other subcategories are almost error-free. This illustrates the validity of our top down approach from suspect category to finer scaled subcategories.

5.6 Missing Data

Another objective of data profiling is to gain information about missing attribute values. An univariate analysis first determines the suspicious attributes. Multivariate analyses are then carried out in different levels of detail for each product type, product category or creation period.

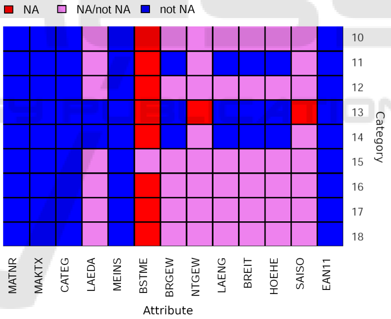

Figure 8: Visualization of attributes with missing values by categories. The colour determines how often values of an attribute for a certain category are missing. Red indicates always missing values, blue always filled values and violet a combination of both missing and filled values. For category 13 all values of attribute NETGW (net weight) are missing for example.

Figure 8, for example, shows missing values in a three-level gradation for each merchandise category. Suspicious categories are then analysed in more detail.

5.7 Reference Products and Locations

In many companies, reference product and location data records are used to simplify the creation of product data. In SAP Retail, for example, a reference product item can be defined for each merchandise category. The reference data contain default values that can be adopted by the user. Using various multivariate analyses, we propose to determine the frequency of deviations from the default values. The reference locations must be taken into account for the logistic data segment. Frequent deviations are suspicious, they can indicate incorrect default values for example.

5.8 Change Documents

In some cases certain attribute values of product items cannot be explained by a multivariate analysis of the current product data record. To support a root cause analysis, we have integrated some functions into our tool that enable an analysis of change documents. A change document describes when, how and with which transaction of an ERP system an attribute value was changed. This should enable a complete analysis of the product data over their life cycle. Currently, among other key figures, the frequency of changes by attribute and product category and the frequency of value changes with respect to a specific attribute are determined. Product categories with many changes must then be subjected to further profiling steps.

6 CONCLUSION

In the digitalized environment of retail data quality is absolutely essential. To ensure a high data quality data profiling has to be performed. In this paper we have presented a reference model and best practices to profile product data in a retail ERP context efficiently and comprehensively. The objective of our approach is to reduce the effort needed to obtain information about the content, structure and quality of product data by taking into account their specific properties and domain knowledge. A first evaluation of the approach was promising, but additional research is necessary.

Moreover, in our opinion the importance of data profiling and exploration is increasing. For instance, many companies are analysing the potential of machine learning to improve their processes. As is well known, machine learning requires training and test data of high quality. If product data are used, our approach can deliver useful information to support a subsequent clean up and labelling of the data for machine learning purposes. Due to the enormous amount of data the cleaning of the product data cannot be done manually, though. Semi-automated cleaning using simple static data rules can only address a part of the problem. Therefore our future research also focuses on the development of intelligent algorithms for product data cleaning itself.

ACKNOWLEDGEMENTS

The authors would like to thank the German Federal Ministry of Education and Research for supporting their work through the KMU-innovativ programme under grant number 01—S18018.

REFERENCES

Abedjan, Z., Golab, L., and Naumann, F. (2015). Profiling relational data: a survey. *The VLDB Journal*, 24(4):557–581.

Di Blas, N., Mazuran, M., Paolini, P., Quintarelli, E., and Tanca, L. (2014). Exploratory computing: a draft manifesto. *Proceedings of the 14th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 577–580.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78.

Guido, A. L., Paiano, R., Pandurino, A., and Pasanisi, S. (2015). Searching issues: a survey on data exploration techniques. *IJETTCS*, 4(6):183–188.

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., and Heer, J. (2012). Profiler: integrated statistical analysis and visualization for data quality assessment. *Proceedings of the International Working Conference on Advanced Visual Interfaces*.

Olson, J. E. (2008). *Data quality: The accuracy dimension*. Morgan Kaufmann.

Papenbrock, T., Bergmann, T., Finke, M., Zwiener, J., and Naumann, F. (2015). Data profiling with metanome. *Proceedings of the VLDB Endowment*, 8(12):1860–1863.