

# Data Science Challenges of Automated Quality Verification Process in Product Data Catalogues

NIEMIR Maciej<sup>1,2,a\*</sup> and MRUGALSKA Beata<sup>2,b</sup>

<sup>1</sup> Łukasiewicz – Poznań Institute of Technology

<sup>2</sup> Poznań University of Technology, Poland

<sup>a</sup>maciej.niemir@pit.lukasiewicz.gov.pl, <sup>b</sup>beata.mrugalska@put.poznan.pl

**Keywords:** Product Catalogues, Product Data Quality Management, Master Data Synchronization, Machine Learning in Data Quality, GPT-3 in Product Catalogues

**Abstract.** Product master data are an essential and key component of purchasing processes, ensuring the smooth running of business operations within companies. Unfortunately, due to the lack of a single, complete, worldwide information system storing reference data, managing the data, maintaining its quality, reliability, and timeliness, requires building quality assurance teams for such processes in most companies. There are numerous errors in product data, and identification and correction of them are time-consuming, especially for large data sets that contain many millions of products. These errors are due to the so-called human factor but are also the result of technical errors and limitations of IT systems. Therefore, in the paper, we proposed a number of solutions by category and group that can automate, simplify, and shorten the master data management process. There are also presented examples of data validation using a variety of techniques, rule-based, dictionary-based, and machine learning, that enable mass verification of both images, textual parameters, digital parameters, and classifiers, while indicating the probability of errors in specific attributes as well as in their combination, and in some cases correcting or proposing correct records. The performed tests illustrate the magnitude of problems and potential on a sample dataset.

## Introduction

The exchange of digital information is an indispensable part of modern commerce. Good-quality and complete data play a key role [1] [2], as they have a significant impact on the efficiency of business transactions, as well as minimizing the occurrence of errors. Negligence can negatively affect the credibility of companies [3] and generates unnecessary costs [4].

One of the most important areas in terms of information exchange is product data, which is extremely difficult to standardize and maintain due to its diversity in terms of properties and parameters depending on the product group. Work on standardization in various aspects and levels of depth has been ongoing for years. The European Commission is working on regulations, in the interests of consumer safety, the GS1 Organization (GS1 GDM - Global Data Model) is developing models, which indicate the most important elements from the point of view of B2B and B2C cooperation. However, there are no complete and respected standards in this area for all manufactured products worldwide [5]. The problem is exacerbated by the fact that manufacturers do not universally make information about their products available in digital form, so supply chain participants often build product data sets themselves [6], using various available sources or manually creating them from scratch. As a result, commonly used product attributes, such as identification number, name, brand, photo, composition, net content [5], are entered in different, often incorrect ways [7], which can be particularly noticeable for fields that allow the entry of descriptive data or in the form of graphical objects. This, in turn, is very important when shopping on the Internet, where the key to search is not product identifiers, but precisely their names,



descriptions, and images [5]. Incorrect data are published on shop websites, search engines, and auction portals, and these are further duplicated and propagated in further databases. As a result of this, manufacturers no longer have control over the actual and reliable description of their products on the Internet, and everyone loses [8].

Alternatively, product data can be synchronized with the catalogues of companies that specialize in obtaining and sharing the correct data, through so-called "data pools". A data pool is a centralized data repository where trading partners (retailers, distributors, or suppliers) can obtain, maintain, and exchange product information in an agreed format. Suppliers can upload data to the data pool, which retailers receive through their data pool. However, not everyone wants to use such a solution, pointing out, among other things, its high cost and lack of flexibility [9]. The solution also does not guarantee the complete assortment needed to populate the database and the set of product data attributes required, and the aggregation of data from different sources again raises the issue of standardization, as can be seen in the results of the study of name similarity in product catalogues [7].

Since all the necessary information is usually found on the label or on an information sheet attached to the product (the manufacturer takes care of these data), as already mentioned, companies often decide to build their own data catalogues, adapting them to their own needs and taking care of their quality on their own, improving the acquired data or creating them from scratch for the IT system. This fact is not surprising given that in e-Commerce, a unique marketing description of a product gives a noticeable competitive advantage ([10] ,[11]), hence the need to combine reliable information from the manufacturer with data created for the shop.

Data collection is an expensive process. Assuming an extremely short period of time - for example, ten minutes to retrieve one product from the warehouse, read the data from the label, verify, and correct the data in the IT system - for a thousand products, the time required to complete the task will be about one month's work for one employee, while with forty thousand, which is the average stock of many B2C companies, it is as much as four years of continuous work. If one were to add to this the need to create one's own product descriptions, the time required to create them would significantly lengthen the entire process, and there is no guarantee that, with such a volume, the data will be entered correctly and that the content of the fields will be consistent and complete. And this is where we touch on the main source of data quality issues. These are:

- error, typos, negligence;
- lack of staff training;
- software errors;
- lack of data standardization.

An important solution to the problem of data entry errors is the implementation of appropriate control procedures [12]. To streamline the process of data control and correction and also to minimize its costs, it is important that this role be taken over as much as possible by the information system. For such purposes, companies use Product Information Management (PIM) software. PIM focusses on the central management of product information to support business processes that involve customer-centric product information [13]. In this classic approach, when developing PIM software, developers and analysts focus on creating advanced data structures, over how information is stored, sometimes on workflows. They are not consulted on technical solutions to improve inter-organizational coordination of master data quality [9]. At the implementation planning stage, the product data found are usually unstructured, erroneous, and reconciling stakeholders and agreeing on the scope of the data is in itself a challenge and requires compromise [14]. Quality control of entries recedes into the background, and the solutions developed generally have only basic data control solutions to:

- Protect the database from critical errors that prevent data storage (limitation of character length or record format);

- Take care of basic data consistency (basic obligatory fields according to the data model, taking care of correct relations in database objects);
- Implement basic data masks (e.g. check digit in the GTIN).

Such a solution usually results in tools that are prone to numerous irregularities on the side of data editing, logic, and field content - the system does not monitor such errors.

### **An In-Depth Study on the Possibilities of Controlling and Improving Product Data**

On the basis of a catalogue with 42 million descriptions of different products collected by 50,000 manufacturers (crowdsourcing), an analysis of the feasibility of additional data validation and quality improvement mechanisms was performed and some of them were investigated. The aim was to find various methods and algorithms that would potentially improve the process of automating data control and quality. In doing so, an innovative assumption was made that validation does not necessarily imply the rejection of the data, but allows for interaction to further validate the data, drawing the user's attention to a potential error. This human-machine interaction provided the opportunity to take a broader view of quality, to include the possibility of prompting, and to validate data that are not 100% certain. Basic industry-independent product attributes [5] were analyzed, enriching them with several additional elements available in the database:

- Basic common attributes: GTIN, Product name, Brand name, Product image, Product classification, Net content with unit;
- Additional common attributes: Product webpage, Marketing description.

### **Results and Discussion**

The preparation of data for analysis should always start with the normalization process in order to standardize the data and prepare them for further verification processes. Skipping this step or reversing the order of operations may result in biases in the performance of other validators, such as the measurement of text field lengths, the examination of duplicates, and the application of dictionary rules. Normalization processes must include, for text fields, conversions to a common character set, and decoding of characters (e.g. HTML entities). For other data, it is necessary to standardize the storage form, e.g., the URL standard (RFC 1738) for web pages and conversions to common numeric formats (text to number conversion, decimal separator conversion). The normalization process itself is usually based on established rules, data converters that do not affect the content of the fields.

Algorithms were run on the database studied, resulting in the correction of up to 9.5% of the products, despite the fact that the system already had basic normalization measures implemented. The analysis showed the need for additional rules, such as the removal of non-printing characters, removal of the transition to a new line in single-line fields, conversion of hard spaces, removal of multiple spaces and tabs, and remnants of underprepared data imports (double quotation marks, quotation marks at the beginning and end of the text, etc.). The prepared, normalized data were further analyzed for the feasibility of using various algorithms and techniques that would be used to build effective data quality validators in two areas:

1. Basic error validation and dependency error validation;
2. Quality validation and suggestions.

For the first area, it was assumed that errors detected would result in the absolute rejection of the data and that the system would not be able to make any changes to the database, since the assumption for this type of validation is to eliminate data that do not comply with standards and logic. The second area includes quality control, which will result in a process in which the user will have to make decisions based on recommendations and/or full communication about the consequences of entering poor quality data.

### **Validators Based on Defined Patterns and Dictionaries**

In order to extend the basic validators implemented in PIM-class information systems, it is useful to review the entire data set by grouping the data of individual attributes to look for specific similar errors to develop common patterns for them. When examining the available dataset, many of such patterns were found that had not been captured by previous validators. For example, the following rule was discovered: If product names do not start with a letter or a number, they are probably incorrect. A test of such a rule by manually verifying 1,000 products confirmed its effectiveness (100%). Likewise, one word meant an error in the product name, as it did not describe the product well enough, taking into account its variants, net content, and brand, and was only a common name describing a type of product, e.g., 'bread.' Another rule is the occurrence of multiple specific characters in the product name (e.g. multiple underscores), which appeared to be 100% indicative of an incorrectly entered product in the database. The number of words in the brand name was also examined, more than four words containing two characters each for the two hundred occurrences tested meant an error in each case. It is worth mentioning that many duplicate names were also found after the data normalization process. This type of validation does not only rely on regular expressions (RegEx) to look for specific patterns in the text. It is also possible to verify the size of the images in terms of their technical parameters such as height, width, and resolution, as well as their compression quality. For web pages, it is possible to verify that the page exists and that it opens correctly. Many validators can also rely on dictionary rules, i.e. specifying prohibited content, e.g. prohibited brand names, prohibited words, and prohibited page content (e.g. page under construction).

The implementation of patterns filled a gap in the validators of the "Basic error validation & dependency error validation" class, giving a large impact on the quality results. For the data set surveyed, 21 million attribute errors were detected in this way. However, these types of validators have significant limitations - the dictionaries need to be updated, e.g. with page templates under construction, with a list of forbidden names, while any validator counting the number of words or regulating the use of specific characters does not guarantee that the data entered will be factually correct, that the product name really is a name and that the image corresponds to it, that the product is correctly classified. Furthermore, it has been noted that when users are informed of a specific problem, e.g. too few characters or too few words, some of them deliberately circumvent the problem by entering additional letters instead of actually correcting the data.

### **Data Scraping & Structured Data on Websites**

Data scraping is the process of extracting specific content from a web page. This makes it possible to obtain and process specific selected extracts from any web page. However, this technique has significant limitations if it is not clear where to look on the page for relevant product attributes. The verification will then be narrowed down to looking for commonalities in product names, brands, and GTINs in page titles or headings to find similarities. However, there is a standard, widely publicized by companies such as Google and Microsoft, that describes how to include metadata on a website so that it is clear where to look for attributes, for example, for product data. The standard is described in detail on Schema.org. The extraction of metadata populated on pages compliant with the schema.org standard for product data can be part of the validation of the product website entered into the PIM system and at the same time part of the validation of other attributes such as name, GTIN, brand, and image. However, it should be noted that schema.org only standardizes metadata, i.e., data about the data, not the content of the data, so the data found are not necessarily a source of reliable information.

The potential for using websites for validation was investigated. From the product database studied, 14447 different domains of websites identified as product pages for 5.8M products were extracted, and then one of the pages in the domain was randomly selected and reviewed. The research found that 15.4% of the domains had the product name stored in the structured data on

the sample pages and 2.88% of them had the GTIN number. This gives the number of potential verifications for the products assigned for these domains: 14.66% for names and 5.81% for GTINs in the entire database.

### **Anomaly Detection**

PIM-class information systems for numeric data usually limit values to the range accepted by the database, without paying attention to the constraints imposed by the logic of the field in question. One reason for this is that it is difficult to specify ranges of values when building an information system, especially when they depend on multiple parameters. For example, completely different ranges can be adopted for the net content of a product in the context of different units and for different product categories. However, if the database is filled with a certain number of products, an analysis can be performed on the basis of existing data, and boundary values can be determined with a certain probability of being a potential error. On this basis, the data can be discarded or, safer in this case, a potential data quality problem can be brought to the attention of the user, i.e. a data quality check in terms of 'quality verification and suggestions'. In the product database studied, the net content of the product was analyzed and errors were detected for more than 11% of the products. A deeper analysis of the issue points to the possibility of additional verification methods using machine learning algorithms [15], which makes it possible to verify data not only for extreme data, but also for those that occur within a range if a given numerical field shows such trends. Net content is an excellent example of this type of behavior, as products in certain categories are sold in a standardized packaging or in a certain standardized number of pieces [16].

### **Semantic Understanding of Textual Attributes**

An important area that can significantly improve the validation of text attributes in the context of product data is the area related to natural language processing, in particular, issues related to semantic understanding. Using the so-called transformer architecture [17], on a properly trained language model, it is possible to achieve very good results related to correct text classification, spam detection, grammar verification, spelling [18], text comparison, as well as translation and transformation of text into other text. Data verification capabilities for data imputation [19] are also indicated, allowing attention to deficiencies in critical data for consumer health, such as allergens in food products [16]. Thus, it gives a whole range of possibilities for data correction which are not available to rule-based validators. Unfortunately, the problem in applying such solutions is usually the need to train the model and then properly train the set for appropriate applications, which is a costly process. However, it turns out that this is not the case for all cases and solutions. At present, for some solutions, pre-trained models available under open licenses are sufficient for the tasks, and the fine-tuning can be based on found data, the so-called dirty data.

An analysis was carried out on the studied dataset, on the basis of which the trained BERT ("Bidirectional Encoder Representations from Transformers") model with a classification head learnt from 100k unverified data found in the database (so-called dirty data) was able to successfully assign a product to a classification (in this case the GS1 GPC - "Global Product Classification"). A match of 96% was achieved, while for 4% the algorithm indicated a different category. Based on 200 manually verified samples:

- 44% the algorithm incorrectly suggested a different category.
- 40% the algorithm indicated a better category than the one found
- 8% of the data could not be correctly classified due to an ambiguous product name.
- 8% of the data fit into more than one category and the assignment issue was disputed.

The result of the test tentatively confirms the feasibility and effectiveness of the selected algorithm, even without additional effort to correctly annotate the data. Differences in the result of the algorithm and the user entering the data would be subject to appropriate interaction, an alert with a hint, which fits into the validation class "Quality verification and suggestions".

In addition to the BERT models, which are effective for specialized tasks, but require some training on the relevant data, one can also use the much more extensive Generative Pre-trained Transformer (GPT-3) models [20], which are distinguished by the fact that for some applications they do not need additional training to get good results. The following examples using ChatGPT (GPT-3.5) without any coaching have demonstrated effectiveness in the context of recognizing the correct layout of brand, name and net content in product names, recognizing the correct units of net content, and suggesting the correct product name from the available attributes.

The product names were invented for the study. There was no in-depth research and comparative testing of BERT vs. GPT-3, nor of the effectiveness of the validators on a larger dataset.

*Table 1. Examples of data validation solutions using GPT-3.5*

| <b>Question</b>   | <b>Answer</b>  |
|---|--|
| The product has the brand name "Bludoo", a net content of 1kg, and is a washing powder. What name should it have if it starts with a brand name?              | "Bludoo Washing Powder 1kg"  |
| Is the full product name 'bludoo' appropriate if the convention of such a name should consist of the brand, the common name, and the net content?             | No, the full product name "Bludoo" is not appropriate if the convention for such a name should consist of the brand, the common name, and the net content. In this case, the full product name should be "Bludoo Washing Powder 1kg" as it includes all three elements, the brand name, the common name, and the net content, which adheres to the convention.   |
| Is the product name 'bread' appropriate if it should consist of the brand, name, and net content?   | No, the product name 'Bread' is not appropriate if it should consist of the brand, name, and net content. Bread is a common name for a food product and does not provide enough information about the product. The name should include the brand, the specific product name and net content, such as "Brand X Whole Wheat Bread 500g" or "Brand Y Sourdough Bread 1 loaf"  |
| Is the product name "MEGA Men's Shirt 1kg" correct if the name should consist of the brand, common name, and net content or number of pieces?                 | No, the product name "MEGA Men's Shirt 1kg" is not correct. The weight of 1kg is not appropriate for a shirt as it is not a weight measure that is commonly used for clothing products. Instead, the name should consist of the brand, common name, and number of pieces. A more appropriate name for this product would be "MEGA Men's Shirt - 1 piece" or "MEGA Men's Shirts - Pack of 3"  |
| Is the product name "MEGA Men's Shirt 1pc" correct if the name should consist of the brand, the common name and the net content or the number of pieces sold? | Yes, the product name "MEGA Men's Shirt 1pc" is correct if the name should consist of the brand, the common name, and the net content or the number of pieces sold. It contains the brand name "MEGA", the common name of the product "Men's Shirt", and the number of pieces sold "1pc". This name is informative and adheres to the convention of providing the necessary information about the product.   |
| I already have a product named MEGA Men's Shirt 1pc in the database, and my database should have unique product names. How do I make it stand out?            | There are a few ways to make a product with a similar name stand out in a database:<br>Add a unique identifier: You could add a unique identifier such as a SKU or product code to the name of the product, e.g. "MEGA Men's Shirt 1pc - SKU 123456"<br>Include specific details: You could include specific details about the product in the name, such as the color, size, or material, e.g. "MEGA Men's Blue Cotton Shirt 1pc"<br>Add a version number: If you have multiple versions of the same product, you could add a version number to the name, e.g. "MEGA Men's Shirt 1pc v2.0"<br>Use different naming conventions: You could use different naming conventions for different products, e.g. "MEGA Men's Shirt 01" instead of "MEGA Men's Shirt 1pc"<br>Use combination of above: you could use a combination of above methods to make it stand out.<br>It is important to note that the most important thing is to make sure that the new name is unique and clearly different from existing names, while still providing the necessary information about the product. |
| Is the product name "Bludoo Wshing Powder 1kg" correct?   | No, the product name "Bludoo Wshing Powder 1kg" is not correct. It contains a spelling mistake "Wshing" instead of "Washing". The correct name should be "Bludoo Washing Powder 1kg". It is important to ensure that the product names are spelled correctly, as it can lead to confusion and inaccuracies when searching for or identifying the product.  |
| Is the product name "Happido the orange juice 1l" correct?  | No, the product name "Happido the orange juice 1l" is not correct. The inclusion of "the" is not necessary and it doesn't add any additional information to the product name. Also, it's not common to include the word "the" when naming a product. The correct name should be "Happido Orange Juice 1l", which is more informative and simple. It's important to use conventions when naming products to make them easily recognizable and searchable.   |

The product description contains the following information: "High-performance pork ham, steamed, made of jointed pieces of meat, with the addition of water. The product contains: pork meat - 76.9%, water, modified starch, salt, thickeners: E407, E415; stabilizers: E331, E452, E450, E451; antioxidant E316, spice extracts (with celery), maltodextrin, citrus fibre, pork collagen protein, flavouring, preservative E250". Is something missing from the description if the product will be sold in the European Union?

The product description you provided contains a lot of information about the ingredients and additives used in the product, but there are a few things that may be missing when the product is sold in the European Union (EU).  
 Allergen information: EU regulations require that products containing certain allergens, such as celery, must be clearly indicated in the ingredients list.  
 Nutritional information: EU regulations require that certain nutritional information, such as energy value, protein, carbohydrates, and fat content, must be provided on the label.  
 Origin: EU regulations require that the country of origin of the ingredients must be indicated on the label.  
 Net weight: EU regulations require that the net weight of the product must be indicated on the label.  
 Storage conditions: EU regulations require that storage conditions must be indicated on the label if they are necessary to ensure the product's safety or quality.  
 It's recommended to check with the relevant authorities in the EU country where the product will be sold to ensure compliance with all the regulations.

The following examples are for illustrative purposes and indicate how the model can be used to validate the data. The following are suggestions for using natural language processing algorithms to create quality validators.

- Detection of whether a product category has been selected correctly. Suggesting the correct category;
- Detection of incorrect product names (spam);
- Detection of similar brands and prevention of duplication. Suggesting the correct brand name;
- Verification of spelling and grammatical errors. Suggesting correct data;
- Suggestion of missing data based on learnt domain knowledge;
- Suggestion of a better product name based on existing attributes;
- Suggestion to change/reformat the data descriptions.

**High-performance Pork Ham**

- Steamed
- Made of jointed cuts of meat
- Water added

Ingredients:

- Pork (76.9%)
- Water
- Modified Starch
- Salt
- Thickeners (E407, E415)
- Stabilizers (E331, E452, E450, E451)
- Antioxidant (E316)
- Spice extracts (with celery)
- Maltodextrin
- Citrus fibre
- Pork collagen protein
- Flavouring
- Preservative (E250)

Allergen information: Contains celery.

Nutritional information per 100g:

- Energy value
- Protein
- Carbohydrates
- Fat

Country of origin: [Indicate the country of origin of the ingredients]

Net weight: [Indicate the net weight of the product]

Storage conditions: [Indicate the storage conditions if necessary to ensure the product's safety or quality]

Please note that the information provided above is based on the ingredients and additives you've mentioned, and it's important to check with the relevant authorities in the EU country where the product will be sold to ensure compliance with all the regulations.

**Fig.1.** Answer to the question "Format the description to make it readable for the buyer".

### Enhancing Product Data Validation with Computer Vision

Validation of product data quality is worth enriching with computer vision techniques. With the help of appropriate algorithms, it is possible to verify the correct placement of the product in the photograph, as well as the color and uniformity of the background, which is important for maintaining appropriate standards [5], [6]. When machine learning models are implemented,

objects can be recognized and assigned to a category, confirming the consistency of the product name and description with the photo. It can be verified whether the photo shows the packaging of a product, one particular product, or several products, or a product in a scene showing an example of use. You can verify in the case of clothing whether it is the clothing itself or a model presenting it, or you can verify that the photo contains only the company logo. Such validations are of great importance from the point of view of sales in the e-Commerce market [6]. Good-quality product images can be used to conclusively verify the compatibility and consistency of a product's core attributes. With optical character recognition techniques, data can be read from photos, and with natural language processing techniques, data can be compared with attributes stored in text, and errors or gaps in data can be found, which can be crucial, for example, when determining the presence of allergens in food products [16]. The effects of such validators will still not be 100% effective in object recognition or text recognition for a long time, especially when considering a large number of product categories that differ in industry specifics, but they can help improve quality, so it is worth implementing them as suggestions in the "Quality verification and suggestions" class of validators.

An experiment was run on detecting background irregularities in photos on the tested data set to observe the scale of the problem and the need for techniques using image processing. The validator indicated a background color error of 10% of all images submitted.

## Conclusions

Maintaining good product data quality is key to minimizing costs in the supply chain. With a large volume of product data, manual data verification is uneconomical and time-consuming, so it makes sense to look for alternatives in the form of suitable automated algorithms implemented in PIM-class systems, especially when the data come from different suppliers or when suppliers themselves enter it into a crowdsourcing model. There are no global standards for most product attributes, which means that data quality problems are not just about simple errors, but are often more complex and widespread. Information systems do not provide such complex and comprehensive algorithms for improving data quality. There are many quality improvement techniques. The process of improving the quality of product data should start with data normalization. This step is crucial as it ensures that all data have a consistent format and structure, making them easier to process and analyze. Simple validation rules and dictionaries can be useful for improving data quality, but they have limitations. These methods only check for basic errors, such as incorrect spelling or formatting, and do not examine the meaning of the content. For a more comprehensive and accurate assessment of data quality, additional validation methods are needed. One such method is the use of natural language processing (NLP) techniques, particularly in the area of transformer-based models. It can help to understand the context and meaning of the data, and detect errors such as inconsistencies or ambiguities in the content. Another valuable tool for improving data quality is computer vision. This technology can be used to validate product images and ensure that they are of the appropriate quality and resolution. This can be particularly important for e-commerce platforms, where product images are the main means of information for customers. Finally, a hybrid combination of multiple validation methods can have a synergistic effect and provide a comprehensive quality assessment of product data. This approach can leverage the strengths of different validation methods and compensate for their limitations, resulting in a more accurate and reliable assessment of data quality.

## References

- [1] M. Cao, Q. Zhang. Supply chain collaboration: Impact on collaborative advantage and firm performance, *J. Oper. Manag.* vol. 29 (2011) 163–180. <https://doi.org/10.1016/j.jom.2010.12.008>
- [2] Y. Hole et al. Service marketing and quality strategies, *Period. Eng. Nat. Sci.* 6 (2018) 182-196. <https://doi.org/10.21533/pen.v6i1.291>



- [3] S.A. Qalati et al. Effects of perceived service quality, website quality, and reputation on purchase intention: The mediating and moderating roles of trust and perceived risk in online shopping, *Cogent Bus. Manag.* 8 (2021) art.1869363.  
<https://doi.org/10.1080/23311975.2020.1869363>
- [4] D. Appelbaum et al. Impact of business analytics and enterprise systems on managerial accounting, *Int. J. Account. Inf. Syst.* 25 (2017) 29-44.  
<https://doi.org/10.1016/j.accinf.2017.03.003>
- [5] M. Niemir, B. Mrugalska. Basic Product Data in E-Commerce: Specifications and Problems of Data Exchange, *Eur. Res. Stud. J.* XXIV (2021) 317-329. <https://doi.org/10.35808/ersj/2735>
- [6] M. Niemir, B. Mrugalska. Product Data Quality in e-Commerce: Key Success Factors and Challenges, In: *Production Management and Process Control*, 36 (2022), AHFE.  
<https://doi.org/10.54941/ahfe1001626>
- [7] M. Niemir, B. Mrugalska. Identifying the cognitive gap in the causes of product name ambiguity in e-commerce, *Logforum* 18 (2022) 357-364.  
<https://doi.org/10.17270/J.LOG.2022.738>
- [8] W.K. Putri, V. Pujani. The influence of system quality, information quality, e-service quality and perceived value on Shopee consumer loyalty in Padang City, *Int. Technol. Manag. Rev.* 8 (2019) 10-15. <https://doi.org/10.2991/itmr.b.190417.002>
- [9] T. Schäffer, D. Stelzer. Assessing tools for coordinating quality of master data in inter-organizational product information sharing, In: *13th Int. Conf. Wirtschaftsinformatik*, February 12-15, 2017, St. Gallen, Switzerland.
- [10] T. Wimmer, M. Scholz. Online Product Descriptions–Boost for your Sales? In: *14<sup>th</sup> Int. Conf. Wirtschaftsinformatik*, February 23-27, 2019, Siegen, Germany.
- [11] J. Mou et al. Impact of product description and involvement on purchase intention in cross-border e-commerce, *Ind. Manag. Data Syst.* 120 (2019) 567-586.  
<https://doi.org/10.1108/IMDS-05-2019-0280>
- [12] A. Haug, J.S. Arlbjørn. Barriers to master data quality, *J. Enterp. Inf. Manag.* 24 (2011) 288-303. <https://doi.org/10.1108/17410391111122862>
- [13] J. Abraham. *Product information management. Theory and practice.* Springer, 2014.  
<https://doi.org/10.1007/978-3-319-04885-7>
- [14] L. Battistello et al. Implementation of product information management systems: Identifying the challenges of the scoping phase, *Comput. Ind.* 133 (2021) art.103533.  
<https://doi.org/10.1016/j.compind.2021.103533>
- [15] L. Poon et al. Unsupervised Anomaly Detection in Data Quality Control, In: *2021 IEEE Int. Conf. Big Data*, Dec. 2021, 2327-2336. <https://doi.org/10.1109/BigData52589.2021.9671672>
- [16] K. Muszyński, M. Niemir, S. Skwarek. Searching for AI Solutions to Improve the Quality of Master Data Affecting Consumer Safety, In: *22<sup>nd</sup> Int. Sci. Conf. Business Logistics in Modern Management*, October 6-7, 2022, Osijek, Croatia, 121-140. [Online]. Viewed: 10-01-2023. Available:  
[http://blmm-conference.com/wp-content/uploads/BLMM2022\\_Conference\\_Proceedings.pdf](http://blmm-conference.com/wp-content/uploads/BLMM2022_Conference_Proceedings.pdf)

- [17] J. Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv, 2019, art. 1810.04805. <https://doi.org/10.48550/arXiv.1810.04805>
- [18] S.M. Jayanthi et al. NeuSpell: A Neural Spelling Correction Toolkit, In: Proc. 2020 Conf. Empirical Methods in Natural Language Processing, online, Oct. 2020, 158-164. <https://doi.org/10.18653/v1/2020.emnlp-demos.21>
- [19] W.-C. Lin, C.-F. Tsai. Missing value imputation: a review and analysis of the literature (2006–2017), *Artif. Intell. Rev.* 53 (2020) 1487-1509. <https://doi.org/10.1007/s10462-019-09709-4>
- [20] B. Ghojogh, A. Ghodsi. Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey, OSF Preprints, 17 Dec. 2020. <https://doi.org/10.31219/osf.io/m6gcn>