SEARCHING FOR AI SOLUTIONS TO IMPROVE THE QUALITY OF MASTER DATA AFFECTING CONSUMER SAFETY

Krzysztof Muszyński
Łukasiewicz Research Network - Poznań Institute of Technology, Poland
E-mail: krzysztof.muszynski@ilim.poznan.pl

Maciej Niemir
Łukasiewicz Research Network - Poznań Institute of Technology; Poznan University of Technology, Poland
E-mail: maciej.niemir@pit.lukasiewicz.gov.pl

Szymon Skwarek
Łukasiewicz Research Network - Poznań Institute of Technology, Poland
E-mail: szymon.skwarek@pit.lukasiewicz.gov.pl

Received: August 4, 2022
Received revised: September 22, 2022
Accepted for publishing: September 23, 2022

Abstract

The quality and completeness of the master data has a direct impact on the accuracy of purchasing processes in supply chains. Today, manufacturers and retail chains have both centralized catalogue solutions and distributed repositories supported by appropriate standards at their disposal. Despite the popularization of digitization of the synchronization processes of data describing products, research conducted around the world indicates basic errors that concern packaging at various levels, from the basic item, through cartons, to pallets. Therefore, incompleteness and unreliability of the data force the parties involved in the processes to remove errors, which leads to a deterioration of the sales economic parameters. However, the master data used in both B2B and B2C relations are not only the identifiers, classifiers, and dimension and weight information, but also a set of information on the composition and content of products, e.g. food products, which may affect the safety of consumers.

Therefore, a detailed verification of the information content provided by suppliers and producers for individual participants in the supply chain is required. Such activities require the work of specialized teams of expert auditors who must deliver a verdict on timeliness, quality, and completeness. The elements of Artificial Intelligence (AI), which can take over most of the controlling activities, are the perfect solution for this role. This paper identifies important factors as the places where important decisions are made regarding the approval or rejection of product/master data.

AI will be an important element of content verification in terms of consumer safety. The role of these mechanisms is particularly important in the context of the sale of food products and cosmetics, i.e. items that come into contact with the human body. Automation of these processes using this methodology and self-learning mechanisms will enable mass checking of entire databases in search of places that do not meet user safety requirements. The implementation of such mechanisms, whether in catalogue systems or distributed systems, will improve the substantive quality of product descriptions, and thus increase their usability and safety and build customer trust in the brands of individual manufacturers.

Key words: AI; Artificial intelligence; Supply chains; Catalogues; Master data synchronization; Security; GDM; Global Data Model;

1. INTRODUCTION

Master Data is a key element in commercial processes between suppliers, wholesalers, and retailers. Correct product information ensures proper identification. Previous research indicates that incorrect input data cause errors in the further part of the sales process (GS1 UK, IBM, 2009). According to this report, even 80% of the master data in the UK are incorrect, despite the use of modern techniques of electronic data exchange. This has a direct impact on sales losses (e.g. £ 300 million over 5 years). This affects the overall cost of handling transactions. That is why it is so important to use appropriate IT tools and standards for data synchronization in supply chains.

One of such standards is GS1 - GDM - Global Data Model, which indicates the most important elements from the point of view of B2B and B2C cooperation and what is needed to close product registration, regardless of location. In addition, it indicates which data should be completed in individual countries. It is important that the standard of the Data Model is independent of the adopted technology, which means Technology Agnostic. This means that it can be used both in typical Excel files and in advanced database systems.

The current practices of sharing and exchanging product data via IT systems have become an indispensable part of the business, proving the high maturity of the chain. However, the problem arises when the data are not fully standardized (Tagliabue, 2021) or when each supply chain partner expects something different (Whitehead et al., 2019). This situation is observed, for example, in e-commerce, where data is used not only for electronic exchange (Niemir & Mrugalska, 2021). Regardless of whether companies use Data Models or exchange information beyond any standards, the problem of their quality remains. It often happens that information is incorrectly compiled or incomplete.

Quality can be controlled. It is done by both the authors of product data, i.e. producers, brand owners, or suppliers, as well as retailers or specialized companies from Data Capturing. The question arises of what to do to reduce the risk of errors or incompleteness of data, especially in those places that are important from the point of view of end-customer security.

The authors of this study point to the importance of using not only organizational and technical environments or standards, but also the mechanisms of AI - Artificial Intelligence. Here, AI is an independent element that supports control activities. Additionally, based on the experience gained, the most important places were defined that should be verified using the above-mentioned techniques to maintain the safety of the final consumer.

Only the combination of the use of product data synchronization standards and AI - Artificial Intelligence techniques in control processes makes it possible to improve data quality. This can have a measurable impact on the cost of servicing the supply chain.

2. METHODOLOGY OF RESEARCH

In the lack of reports from the literature related to the use of AI in the field of master data quality improvement (specially customer safety), the focus was on determining the method of evaluating product data for goods in modern supply chains and methods that would enable the automation of the data quality control process.

Its authors decided to divide the research work into three aspects:

- Theoretical, identifying the problem of AI and the possibilities of its application in logistics processes,
- Previous business experience in the field of applied standards such as GDM - Global Data Model in the field of proper shaping of product data and identification of places that would require additional control,
- Practical, presenting the possibilities of using solutions treated as a forward-looking proposition of AI implementation in the control of data crucial for consumer safety.

The research process presented in the article results from the logic of the structural analysis of the identified research problem. The adopted research methodology assumes theoretical research, including the identification of similar, but not the same AI solutions.

According to the authors, the above-mentioned aspects are intended not only to confront research, concepts and practical considerations, but also to organize the existing knowledge on the analysed topic. Ultimately, the specificity of the research problem requires comprehensive research at every level and carrying out the so-called proof of concept. This applies in particular to those algorithms that have been indicated as recommended for the described applications, i.e. improving master data quality control.

3. LITERATURE REVIEW

3.1. The importance of product data

Modern trade requires participants to efficiently use product data to allow proper identification of the subject of sale. Today, the easier it is to promote a product, the more knowledge the market has about it. This confirms the theory that a product is treated as incomplete without adequate information.

There are many definitions of master data in the literature. For example, (Schäffer & Stelzer 2017) defined it as a set of representing data. Such a data set can describe different types of product features or properties - both physical, structural, and compositional (Schäffer & Stelzer, 2017; Niemir & Mrugalska, 2021).

Additionally, product data can be divided into areas such as identification, classification, and description (Legner & Schemm, 2008; Vandic et al., 2018). Undoubtedly, the key attributes of the product data are:

- Product name that uniquely and fully identifies a specific product, taking into account its variant and brand, without having to know other product attributes and without having to browse photos or physically view the product,
- Unique identifier - unambiguously representing the product in the supply chain, created in one standard and interpreted in the same way by all IT systems. A well-known and widely distributed number that meets these requirements is the GTIN (Global Trade Item Number – late EAN Code) issued by the GS1 organization.

Before starting cooperation with wholesalers or retail chains, producers and suppliers prepare the so-called product cards. These are still very often Excel files. Moreover, in parallel, data recipients use independent catalogues or their own data collection tools.

Ideally, the market would expect free access to structured information prepared similarly for the entire market. Access to such information should be available not only to the authors of the content, i.e. producers, brand owners or suppliers, but also intermediaries, wholesalers, retailers, logistics operators, and consumers. This would make it possible to maintain adequate homogeneity and consistency of these data in all the above-mentioned market participants, while guaranteeing unambiguous interpretation. Hence the search for organizational and technical solutions that would enable synchronization of data describing products.

In connection with the above, catalogue systems such as GDSN classes - Global Data Synchronization Network, which is treated as central databases. GDSN is a team of global cooperating product catalogues that meet the requirements of GS1 standards. Thanks to this, the data can be transparently transferred from the supplier to the retailers, regardless of who uses which GDSN catalogue. Additionally, there are also distributed environments based on individual internet resources of individual brand owners (Osmólski & Muszyński, 2020).

Another element in the fight for data quality, are standards, the purpose of which is to unify and simplify the rules for collecting, processing, exposing and transferring the above-mentioned data. They also shorten the processes of goods exchange by eliminating errors related to the correct identification of products. Consequently, it improves the economic parameters of sales.

Since trading occurs across multiple channels simultaneously, product information is needed to support all key supply chain processes. Product information, also known as master data, is also the foundation of proper product identification. This is especially important in the era of the omnichannel, in which all commodity participants should simultaneously have access to correct and high-quality data that describe products. Here, Omnichannel should be understood as sales carried out with the use of integrated content coordinated and available everywhere (classic stores, online stores, mobile applications, social media, and AR - Augmented Reality).

To identify the information scopes needed to support data exchange, the authors of the study divided the key attributes into sections. The most important of them are presented in Table 1.

Table 1. List of the most important groups of attributes describing FMCG products (Fast Moving Consumer Goods = Food and Near Food products)

Attribute sections	Sample content
Supplier and product / package identification	Vendor Identifier - GLN - Global Location Number Product identifier - EAN / GTIN - Global Trade Item Number
Product name and description	Product name / Short name on the receipt
Marketing information	Marketing communication regarding the product
Legal requirements	Regulated product name; Is the price indicated on the product?
Tax classifications and rates	Product classification, eg. - General Classification - GPC; CN code; VAT rate
Information based on Regulation 1169 / EU	Nutritional information; Allergen content; Breeding place; The place of fish catch
The origin of the product	Country of manufacture of the product
Dimension and weight information	Height / width / depth; Net weight; Net content; Weight after draining
Product durability / warranty storage information	Use by / use by date Maximum and minimum storage temperature

Package type	Unit package / Carton / Pallet
Information on palletizing	Number of pieces on a layer / Gross height of the pallet
Packaging material	Cardboard; Foil; Glass bottle
Safety information	Dangerous product marking; The presence of batteries in the product
Complementary attributes for controlling the data publication process	Date of data publication; Is the unit intended for shipping?; Is it an invoiced unit?

Source: own study

In the case of selling food products, the manufacturer should provide information on the composition with particular emphasis on substances that may be dangerous to the health or life of the consumer (e.g. allergens by the EU Regulation - EU 1169/2011). However, when selling through the Internet, it is important to meet the requirements of the Consumer Directive - EU 83/2011. The buyer/customer may return the goods if he deems that the description was too limited and was thus misled by the seller.

Therefore, as you can see, the completeness and quality of product data have a significant impact not only on building brand loyalty, but also on customer safety.

The global standardization, organization GS1 conducts intensive initiatives aimed at structuring product data, which in turn has an impact on the improvement of the quality of information flowing between all participants of the supply chains. One of them is the implementation of the GDM Global Data Model. This organizes the sets of attributes applicable to particular industries. It is also a starting point for the implementation of technical platforms and other standards related to the quality of information about products (Muszyński, 2021).

Today, suppliers, retail chains, Logistics Operators and Catalogue Providers associated with international GS1 Working Groups have developed Data Models that accurately describe data in the following sectors/industries:

- Food,
- NearFood,
- PetFood,
- Alchohol&Beverages,
- Tobacco,
- From 07/2022 also DIY – Do It Yourself.

It should be clarified here that the description of typical FMCG products is a result of at least two Food and NearFood Models.

GDM defines not only what the scope of information should be broken down by industry, but also indicates which of the attributes are required. This is the important factor, which additionally indicates which of the attributes are absolutely mandatory. It means that without this it will not be possible to properly close the registration of the product in the supplier-recipient relationship (e.g. Retailer).

Regardless of the described standards and EU regulations, a business carries out its sales mission in its own individual way. Important elements used in B2B and B2C relations are, first of all, the connection of the GTIN identifier (Global Trade Item Number) with the name of the product, then with its detailed description, and in the case of e-commerce with a photo.

Data catalogues often use validation rules that allow you to catch basic errors such as:

- Incorrect GTIN syntax - wrong GTIN number/duplicate existing GTIN,
- Arithmetic errors, e.g. incorrect number of unit packages in the box,
- Logical errors - wrong assignment of product categories.

However, the above-mentioned solutions do not provide certainty of entering data of appropriate quality in all cases. Hence the need for a deep connection of several factors. And here comes the place to implement AI methods to better control the shared data.

Examples of problems related to the identification and interconnection of relations between the attributes are presented below.

In the example cited above, in traditional trade, the key attribute of a product is GTIN, which uniquely identifies the product, while in e-commerce, product identification begins with its name (Niemir & Mrugalska, 2022). For example, in a stationary store, the customer identifies the product by its packaging, and the purchasing process begins with scanning the bar code - that is, GTIN identification, the IT system uses it to determine the price, inventory, etc. In the case of online purchases - the product is identified and searched for by name. At this stage, the GTIN does not matter and the purchase is made at the level of the internal online store ID associated with the name. This shows why the product name is so important. Unfortunately, it turns out that just as the GTIN number is a permanent element, given by the manufacturer, the name of the product is interpreted by everyone in their way.

Using one of the local purchasing platforms, research was conducted on a product with the same GTIN number entered for many names of offers and categories. This is a classic example where a buyer searches for different products and consequently receives multiple offers with the same product. Platforms are currently trying to solve such problems by creating their own product catalogues, and aggregating such data by GTIN.

This is only possible if the platform has reference data. This solution is not 100% effective. In the discussed case, as many as 771 pairs were created for the set: offer name (product) + product category. The product appeared in 10 different categories and 283 different names. Instead of the name of the product, the offer contained various information - from the common name to the description of the client's needs. This demonstrates the seller's approach of using an offer description to attract attention at the expense of sound product information, only to increase sales. This is an example of the so-called wrong combination of attributes.

Another multi-criteria problem may appear when we match the name of the product with its photo. The study analysed all stores in detail to confirm the selected word about the name, to eliminate situations when the search engine retrieves data, e.g. from the description on the product page.

There are many reasons for this, such as an incorrectly entered product name by the operator of an on-line store database or a badly loaded photo that does not represent the real product, or even a poor-quality photo, e.g. a photo, distorted colours, the wrong colour definition.

Another problem is the photo itself - the shot, background, number of products in one photo etc. In the traditional channel, such situations are extremely rare, as the consumer can physically touch the product and verify his own experiences/feelings. Given the above examples, it seems that, AI algorithms and methods will significantly improve the data entry process and eliminate discrete errors that cannot be detected using the classic validation methods. The above reports prove that, as a user, we have some data, but they are "dirty". In addition despite we use of standards, electronic catalogues, etc. The numbers cited above, as well as the types of problems, and the fact that standard guidelines, validators, best practices, and software vendors' security are far too high-quality data. By analysing the literature on the subject, we found no relevant publications on how to improve the quality of product catalogues. Therefore, we decided to analyse the publications in terms of AI algorithms available to improve the quality of the data.

4. DISCUSSION - LOOKING FOR A SOLUTION

In this chapter, the authors try to select information areas to be controlled and the best potential algorithms currently available on the market. The work does not examine their details, but only indicates the potential in their application.

4.1. Description of the problem to be solved

From the point of view of product data security, the following issues should be taken into account, in particular:

- Key attributes determining consumer safety,
- Current and future data verification methods.

Among the attributes presented in this study, the authors rejected those that do not have a direct impact on consumer safety. These are, among others, product identifiers, manufacturer and product names, and dimension and weight information. These data are important mainly for handling B2B relations between the supplier and the retailer. There remains a group of attributes related to the characteristics of the product, such as composition, nutritional requirements, suitability, and information on hazardous materials.

Finally, from among the group of attributes, the most error-prone were selected. It is behind them that the safety of consumers stands. These are:

- Classification of the product,
- Name of the product,
- Regulated product name,
- Functional name,
- Product description (details of the product),
- Does the product contain substances dangerous to health and life,
- The content of allergens,
- Maximum date for consumption,
- Nutritional information / calorific value,
- Nutritional information / content of protein, sugar salts, etc.,
- Allergen warning on the product packaging.

When using electronic master data synchronization methods, the data pools often use simple validation rules. They consist of checking the arithmetic values and the logical presence or absence of some descriptive fields.

As it turns out, such a cursory check is not always sufficient. A content check is required. It is most often done by a man who verifies and allows the description of the data for further use. This means that the data will be in a retailer or an online store. The human factor is the possibility of a mistake. In addition, there is an effect of scale. How to efficiently control the base of 40 thousand items (typical index list amount in retailer)? What about the situation at the data pool, when there are millions of GTIN-s. Here, self-learning methods derived from AI algorithms are needed. Therefore, it is suggested to use Artificial Intelligence mechanisms for faster and more efficient identification of defective data in order to release for use only those that meet the control conditions.

4.2. Indication of AI algorithms and methods needed to verify product data

Focusing on consumer safety, a group of the most important attributes was selected. They also form the core of the information defined in Regulation 1169 / EU. They include: Product names, Detailed descriptions, Classifications, Information on hazardous material content, Nutritional information, Shelf life, and Allergen warnings. The literature in the area of broad Artificial Intelligence supporting data analysis including text and images was then analysed in search of possible algorithms that could improve the quality verification of these attributes. As a result, several ideas for verification were selected that potentially offer a chance for effective implementation of solutions of this type, as presented in Table 2 in the combination: idea - detailed description.

Table 2. AI and detailed information on selected attributes important for consumer safety

Idea	Description
Correctness of product classification in the context of the name and description	A model learned from the data could determine the correct classification of the product. If the classification differed significantly from that assigned to the product, it could be reported as a data error.
Cross-validation of name and description in the context of safety and allergen attributes, key characteristics, and net content	Algorithms in the area of text transformation and natural language processing could extract data recorded in descriptive form, such as in the name or marketing description, and then compare it with data contained in the form of lists or yes/no information. Deficiencies or differences could be reported as errors.
Missing ingredient and allergen data	Assuming the product dataset is sufficiently large, it could be possible to predict missing elements of multi-element attributes (e.g., ingredient or allergen lists), non-matching data, or non-existent data.
Incorrect numeric data in attributes	In large product datasets, it is possible to observe some recurring rules between product attributes in the same categories. Using appropriate machine learning algorithms, deviations from the norm could be detected.
Incompatibility of product image with text attributes	Using techniques such as computer vision (CV), machine learning (ML), optical character recognition (OCR), and natural language processing (NLP), it is possible to effectively extract data from a product label, classify it, and compare it with data stored in individual attributes.

Source: own study

4.3. Correctness of product classification in the context of the name and description

Correct product classification is significant for consumer safety. With proper product classification, product management systems, sales platforms and other IT solutions can require additional fields dedicated to certain product groups, limit accessibility for underage consumers, etc. The classification problem is generally known, and research in this area on the use of Artificial Intelligence has been carried out for many years with good results (Lee et al., 2021). However, the authors propose a different approach to this topic, the goal of which would be not to find the correct classification of the products in question, but to find an error in the product data that has already been described and return the information to the data controller. The benefit of such a solution would be a reduction in data preparation costs. This solution could be used to improve the quality of data classified using GS1 GPC (Global Product Classification), as well as any other classifications, provided the data used to learn the algorithm is large enough and sufficiently fills each branch of the classification.

The choice of model for text classification is not a trivial task. There are many classical machine learning approaches for classification (Kowsari et al., 2019). The basis for the operation of any such algorithms, is first to properly convert the words of the input text into corresponding numbers or multidimensional numeric vectors, the so-called "word embedding". Classical embedding methods, such as TF-IDF ("Term Frequency - Inverse Document Frequency"), work by counting words and assigning corresponding weights to them, which, however, has major drawbacks - the order of words in the text is ignored, the algorithms are unable to capture semantic information, and they have high dimensionality. This limits their application to effective text classification tasks (Naseem et al., 2020). Novel, but more computationally expensive approaches are based on transformer architectures. They use a pre-trained encoder on large text datasets, e.g., BERT ("Bidirectional Encoder Representations from Transformers") enhanced with a classification head (Devlin et al., 2019). Studies show that such solutions are significantly better than classical methods (Gasparetto et al., 2022).

Text classifiers require training. Training of such an algorithm is usually carried out on correctly prepared labelled data, which is prepared by a trained team of so-called annotators. If the classifier is large, hundreds of thousands of data need to be annotated, which may not be cost-effective in relation to the business benefits. However, if one were to run the learning of the classifier on real, dirty data and perform data training using stratified cross-validation (Bates et al., 2022), which should minimize the adverse impact of errors on the model, there is a chance that with a large scale of data the results will be sufficient enough to achieve the desired goal. This is because the solution is not about the correctness of the classification prediction, but about alerting the database administrator to a possible error present in the data. Success, therefore, would already be the narrowing of the search for errors by the database supervisor to, for example, 10% of the entire database, which still saves 90% of the time. Of this 10%, some of it would certainly be an error in the operation of the algorithm, whereas a properly set up information system could retrain the model by using the information resulting from the administrator's final corrections.

4.4. Cross-validation of name and description in the context of safety and allergen attributes, key characteristics, and net content

Computer systems usually do not validate the content of fields if they contain free text. As a result, data discrepancies are observed between product data attributes. Various algorithms from the field of natural language processing (NLP) trained and tailored to the particular language in which product data are stored could help extract relevant and useful information for comparison with other product attributes (Srinivasa-Desikan, 2018). In this way, information about the content of a given ingredient (e.g., active ingredient, allergen, hazardous substance) or the net content of a product could be extracted from product names stored in the database (Javed Akhtar et al., 2020; Zheng et al., 2018). This information could be used to verify that the name matches the actual data recorded in detailed product attributes. This concept applies to various product names: the functional name - describing how the product is used by the consumer, the regulated name – it means the recommended, regulated or generic name or term of the product describing its true nature. This is all, the basic field, which is the product name (Niemir & Mrugalska, 2022), sometimes referred to as the label description, as a consumer-friendly short description of the product suitable for compact presentation. Similarly, a product description, regardless of its form (e.g., a marketing description), should contain full information about the product, including restrictions on use due to allergens or hazardous substances, which various data extraction methods can be used to control (Wang et al., 2020a;, Wang et al., 2020b). On the other hand, if the product information database contains yes/no information, and extracted lists containing ingredients, allergens and hazardous substances - these completely must coincide with the description. Describing the issue more broadly - Artificial Intelligence by cross-referencing product textual data such as name, product description, lists with ingredient and allergen information could verify their correlation and automatically detect potential anomalies, and report these to the data manager.

Data consistency is very important for consumer safety. Often the name, description, and detailed attributes of a product come from different sources, are incompletely aggregated, or entered by other people. While inconsistencies in numerical values between a product's name and its description or detailed parameters can be mainly a cause of consumer confusion, already false, unconfirmed, or contradictory textual information can lead to serious health consequences. For example, when a product whose name includes the statement "gluten-free" will have the information "may contain gluten" in the description and no warnings in the allergen fields, while in fact it actually contains such an ingredient.

The subject area of this type of field validation is very broad and can include many effective solutions worth testing. One of the simpler methods here may be the extraction of net content from the product name. With the use of appropriate regular expressions and a well-described dictionary of entities, along with synonyms and abbreviations, the effectiveness of such a solution can be assumed to be high. As a result of the operation, missing extracted data, inconsistent units or values, would be a reason to report an error to the data manager. A parallel approach could be a more high-tech algorithm that examines text consistency at the level of specific words, programmed to recognize standard phrases - that is, the basic context of use, such as: "contains [...]", "does not contain [...]", "may contain traces of [...]", and other cross-contamination problems. Using appropriate methods that perform lemmatization, i.e. reducing each word in the text to its base form, and a programmed dictionary of synonyms (Latin descriptions, regulated food additives described as E, etc.). The solution could also be characterized by high efficiency. In this case, the data manager could be informed of both missing information and inconsistencies, especially with the phrases: does it contain or does not contain.

The most high-tech solution worth testing and promising good results would be, as in the previous consideration, to perform calculations on text data after converting them into vectors of numbers. There are a number of embedding methods, such as Glove Embedding - but the method does not encode the context information that is crucial here, the ELMO method ("Embeddings from Language Models"). This encodes bidirectional context information (Peters et al., 2018), and finally the most interesting and noteworthy - the aforementioned BERT method (Devlin et al., 2019), which best encodes context using the attention mechanism (Vaswani et al., 2017). Several calculations can be performed on such a transformed dataset: calculating the similarity of a given, e.g., allergen to a single word in a text using Cosine Similarity (Gomaa & Fahmy, 2013). It gives similarly to the aforementioned method with a classification head - to determine whether, e.g., a product contains allergens, information about warnings, usage restrictions, etc. or not. A slight similarity of a word, e.g. allergen to a product description, or a different response from the system in the category of yes/no fields would result in sending the information to the data manager. Unfortunately, such algorithms require training on tagged data. The better the quality of the data - the better the final results of prediction effectiveness, although, again, tests can be performed on dirty data knowing the consequences of shifting the training process to the data admin in the initial period of operation of such a system.

4.5. Missing ingredient and allergen data

Another problem of data quality is missing parts of the information. If the data in the product catalogue do not have extensive descriptive information including, for example, product ingredients or allergens, and the list with ingredients or allergens is not complete either, the cross-validation algorithms described in the above paragraph will not work. When verifying such information, an experienced data manager is likely to point this out, but due to the large amount of data or inexperienced staff, the fact of missing data may be bypassed. It turns out that the process can be automated. Assuming the product dataset is sufficiently comprehensive, it is possible to predict missing elements of multi-element attributes (e.g., ingredient or allergen lists). This is known as data imputation (Lin & Tsai, 2020). In this case, the data would not automatically be entered into the database, but information about possible deficiencies would be forwarded to the data manager for verification. In this way, it would be possible to alert the data manager to a possible oversight in product data, potentially reducing the risk of a missed issue. This is particularly important where a lack of information could have consequences in terms of endangering the life or health of the consumer.

One of the preferred solutions to the problem could again be the use of text classification based on BERT transformers, as described in the first issue ("Correctness of product classification in the context of name and description"). In this case, the classifier would be, for example, a list of ingredients and allergens created using the One Hot Encoder technique (García et al., 2016). A text classification would then be performed, with the difference that on top of the architecture would be a multi-classification head with the ability to select more than one class for a single product, since a product may contain more than one allergen or ingredient. The result of such an algorithm would be a list of likely items that should be listed, in terms of percentages, which, after comparison with the actual data, could be sorted and evaluated by the data manager.

4.6. Incorrect numeric data in attributes

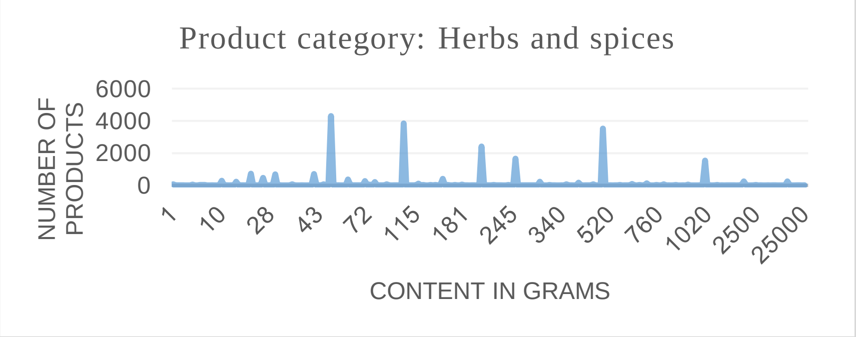
In product databases, one often encounters anomalies in numerical data that "meets the eye." This is the kind of data that the information system records because the value is within a preset range, but a human can identify it as "illogical" at a glance. These are the results of not only human mistakes in terms of figures or units but also wrong data conversions between IT systems. Although some attributes, such as those related to logistics, do not directly affect consumer safety, errors in some data can have more serious effects. Consumer safety is affected, for example:

- Allergen content - the value / amount of the allergen,
- Maximum date of consumption / use from delivery,
- Maximum date of consumption / use from production,
- Nutritional information / caloric value,
- Nutritional information / content of protein, salt sugar, etc.

When defining extreme numerical values of given parameters, values can be estimated from experiments, but this is a time-consuming operation given that the permissible parametric ranges also depend on units, as in the case of net product content. It is much easier to use algorithms from the field of statistics and perform calculations on existing data sets. The easiest way - you can calculate minimum and maximum, or, with a large dataset, the median from the set of all minimum, and maximum values on the products, and thus determine threshold values, beyond which the data manager will get information about a potential error. Such a solution will reduce anomalies, but will also introduce additional false positive errors on extreme values incorrectly identified by the system as potentially erroneous.

Analysing the issue in more depth, the ranges of correct figures can vary from one type of product to another. Moreover, on a large dataset of packaged products, some recurring rules were also observed depending on the classification to which the product was assigned.

Figure 1. Number of products / net content in the category of herbs and spices



Source: own study

9Figure 1. shows the specific sizes of the number of products (from different companies) for a specific value of the number of grams. It can be seen that certain values are preferred by manufacturers. This, in turn, gives an introduction to consider the possibility of using advanced machine learning algorithms in anomaly analysis (Poon et al., 2021) and detecting not only extreme numerical values, but also data that occur inside the range. It is possible to detect such correlations and draw the attention of the data manager to deviations from the norm. Unfortunately, the quality of such prediction depends on the already accumulated data in the database.

There are various models that can be used to detect anomalies. Poon et.al. (2021) list as many as 11 of them in their research on this type of detection, without identifying the best one. It is worth mentioning one of the simpler and older K-Nearest Neighbours models - KNN (Guo et al., 2003) which stores all available cases and classifies new ones based on a similarity measure, an algorithm using density level estimation based on a threshold of the number of neighbours - Density-Based Spatial Clustering of Applications with Noise - DBSCAN (Hahsler et al., 2019) and Isolation Forest, which identifies anomalies using a decision tree algorithm (Hariri et al., 2021). All should be verified at the case study stage.

4.7. Incompatibility of product image with text attributes

In the last idea discussed, it is worth considering the possibility of cross-validation between product packaging images and the product description and other text attributes. This could be the ultimate validation of data quality, especially important given that in large corporations, the responsibility for product label design and country compliance undergoes a completely different quality control than the process of publishing textual data. What is more - you can be sure that the data placed on a product photo come from the manufacturer, while text data have no such guarantee. Verification of data using images is fundamentally dependent on their quality, including resolution and coding, and cannot be carried out if the product packaging images do not depict the product label - which often requires analysis of several images, such as each side of the carton separately. However, there is a big improvement in both the quality and completeness of images, due in part to e-commerce requirements. A photo-description cross-analysis solution could at the same time verify deficiencies in textual data, but also examine whether the posted photo is appropriate and corresponds to the described product, and thus unequivocally assess the consistency and reliability of the data, taking into account the data determining product safety.

The process of extracting information from product images would have to start with computer vision (CV) issues, i.e. analysing the image and extracting blocks of background-free text (Gundimeda et al., 2019). One of the more effective usable solutions could be the EAST model - "Efficient accurate scene text detector" (Zhou et al., 2017), one of the best for detecting text in natural scenes. Then, a ready-made, learned OCR (Optical Character Recognition) text recognition tool, such as "Tesseract Open Source OCR Engine" (Smith, 2007) could be used to extract text in previously detected text blocks from product images. The extracted text, could be used for semantic similarity analyses using BERT transforms and cosine similarity, as presented in the idea "Cross-validation of name and description in the context of safety and allergen attributes, key characteristics and net content ". In this way, the collated and compared data would give full information about the correlation of the photo with the description, and the completeness of both. Any discrepancy would be grounds to alert the data manager to potential quality problems.

5. CONCLUSION

The constant emphasis on the use of digital media, and the need to satisfy the specific appetite for data on the part of informed consumers are the basic determinants defining the current trends in the field of product information exchange/synchronisation. Therefore, you can see that there is no turning back from the current practice.

The observations and experiences of the authors described in this study allow us to conclude that to maintain the high quality of the product data, it is absolutely necessary to implement two mechanisms in all market participants:

1. Organizing product information, eg. by the GDM standard - Global Data Model. This standard will ensure not only the systematization of data, but most of all it will affect the presence of key attributes, i.e. it will provide a complete set of data. Here, all information critical to maintaining consumer safety is particularly important.
2. Implementation of control procedures for data collected electronically, using AI. It will allow to eliminate of many errors related to incorrectly compiled information or even the lack of some data. This should be seen as automating data verification procedures. Until now, verification was done manually or with the use of simple validation rules.

Global Data Models unify and systematize the information scopes necessary to conduct efficient sales. Thanks to them, all trade participants have sufficient knowledge about what data are globally important and what are important to meet the legal requirements of the European Union. The implementation of the aforementioned order in the form of a Data Model translates into an improvement in the cost parameters of several activities related to the preparation and transfer of product data. Information supplemented according to this pattern enables full control over their quality. This, in turn, is related to the improvement of further parameters such as consumer safety or creating a consistent policy of trust in the brand.

The concepts mentioned in the study are the starting point for building product data exchange systems that will independently take care of storing high-quality data, i.e. building modern and trusted sources of information about products.

To sum up, the group of attributes regarding e.g. composition, content and durability of the product should be controlled not only manually, but with the use of selected AI methods/algorithms. The review of the recommended methods and algorithms is described in detail in the previous chapter. This guarantees the improvement of data quality and, consequently, the possibility of influencing several key parameters:

- Consumer safety,
- Brand trust,
- Eliminating errors and improving the economics of sales.

By extrapolating the conclusions described in this study, Artificial Inteligence can also be ultimately used for the processes of automatic information improvement. Thus, AI deep learning systems would enable the completion and refinement of data. This, in turn, fits in with the concept/trend of an ideal market in which all its participants would have access to the same qualitative information describing products.

This article indicates the areas for optimizing the quality of product data, i.e. a group of key attributes for customer safety. Recommendations regarding the selection of AI methods and algorithms were also presented. All the elements gathered above are theoretical considerations. Now it is necessary to carry out the implementation and test whether the described indications are sufficient to meet the requirements of full quality control of master data or not. Therefore, it is suggested to carry out the proof of concept on a sufficiently large set of real data. Real data, means, obtained from the market (B2B and e-commerce data). This will certainly give answers to whether the selected AI methods are closely related to the type of base, or whether methods and experiments can be easily transferred to any set.

In the initial phase, each of the methods/algorithms requires manual control of the test moderator, i.e. the process annotator. Only in this way will it be possible to evaluate the effectiveness of selected solutions, and it will also be possible to carry out tests of AI resistance to disruptions in the processes of improving/enriching of product data.

The last task that should be carried out as part of the proof of concept is to test the profitability of the implementation. That is, indicating the optimal volume of the data set for which it is worth paying the cost of implementing AI technology and the personnel cost of the implementation team. Only such a set of information will give the data pool provider a complete picture of what AI algorithms are to be used, how to do it and when to do it. It means, at what point in the development of the database system.

6. REFERENCES

Bates, S., Hastie, T. and Tibshirani, R. (2022). Cross-validation: what does it estimate and how well does it do it? arXiv. Available at: http://arxiv.org/abs/2104.00673 (Accessed: 23 August 2022).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv. Available at: http://arxiv.org/abs/1810.04805 (Accessed: 23 August 2022).

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 1-22.

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A Survey on Text Classification Algorithms: From Text to Predictions. Information, 13(2), 83.

Gomaa, W. H., & Fahmy, A. A. (2013). Article: A Survey of Text Similarity Approaches. International Journal of Computer Applications, 68(13), 13–18.

GS1 UK, IBM (2009). Data crunch report: The impact of bad data on profits and customer service in the UK grocery industry, The Institute of Grocery Distribution, Cranfield School of Management.

Gundimeda V., Murali RS., Joseph R., Naresh Babu NT. (2019). An Automated Computer Vision System for Extraction of Retail Food Product Metadata. In: Bapi RS, Rao KS, Prasad MVNK, eds. First International Conference on Artificial Intelligence and Cognitive Computing. Advances in Intelligent Systems and Computing. Springer; 199-216. doi:10.1007/978-981-13-1580-0_201,

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). Springer, Berlin, Heidelberg.

Hahsler, M., Piekenbrock, M. and Doran, D. (2019) dbscan: Fast Density-Based Clustering with R, Journal of Statistical Software, 91, pp. 1–30. Available at: https://doi.org/10.18637/jss.v091.i01.

Hariri, S., Kind, M.C. and Brunner, R.J. (2021) 'Extended Isolation Forest', IEEE Transactions on Knowledge and Data Engineering, 33(4), pp. 1479–1489. Available at: https://doi.org/10.1109/TKDE.2019.2947676.

Javed Akhtar M., Ahmad Z., Amin R. (2020). An Efficient Mechanism for Product Data Extraction from E-Commerce Websites. Computers, Materials & Continua. 65(3):2639-2663. doi:10.32604/cmc.2020.011485

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

Lee N., Kim J., Shim J. (2021). Empirical Study on Analyzing Training Data for CNN-based Product Classification Deep Learning Model. The Journal of Society for e-Business Studies. 26(1):107-126. doi:10.7838/jsebs.2021.26.1.107,

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), 120-152, Lin W.C. & Tsai C.F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev. 53(2):1487-1509. doi:10.1007/s10462-019-09709-4,

Muszyński K. (2021). Nowoczesna wymiana danych produktowych, w: Koliński A., Stajniak M. (red.), Zarządzanie i optymalizacja procesów logistycznych we współczesnych trendach gospodarczych, Instytut Naukowo-Wydawniczy „Spatium", Radom, 65-80 – 2021.

Naseem, U., Razzak, I., Khan, S. K., & Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. Transactions on Asian and Low-Resource Language Information Processing, 20(5), 1-35.

Niemir M. & Mrugalska B. (2021). Basic Product Data in E-Commerce: Specifications and Problems of Data Exchange. ERSJ. XXIV(Special Issue 5):317-329. doi:10.35808/ersj/2735

Niemir M. & Mrugalska B. (2022). Product Data Quality in E-Commerce: Key Success Factors and Challenges, Production Management and Process Control, Vol. 36, 1–12 .

Osmolski, W., & Muszynsk, K. (2020). Monitoring Of Goods-Documentation Flows In Modern Logistic Supply Chain, Based On Blockchain Technology. Business Logistics in Modern Management.

Peters, M.E. et al. (2018) 'Deep contextualized word representations'. arXiv. Available at: http://arxiv.org/abs/1802.05365 (Accessed: 15 September 2022).

Poon L., Farshidi .S, Li N., Zhao Z. (2021). Unsupervised Anomaly Detection in Data Quality Control. In: 2021 IEEE International Conference on Big Data (Big Data). ; 2327-2336. doi:10.1109/BigData52589.2021.9671672.

Schäffer, T., Stelzer, D. (2017). Assessing tools for coordinating quality of master data in interorganizational product information sharing. In WI2017.

Smith, R. (2007) 'An Overview of the Tesseract OCR Engine', in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), pp. 629–633. Available at: https://doi.org/10.1109/ICDAR.2007.4376991.

Srinivasa-Desikan B. (2018). Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, SpaCy, and Keras. Packt Publishing Ltd.

Tagliabue, J. (2021). You do not need a bigger boat: recommendations at reasonable scale in a (mostly) serverless and open stack. In Fifteenth ACM Conference on Recommender Systems, 598-600.

Vandic, D., Frasincar, F., Kaymak, U. 2018. A framework for product description classification in e-commerce. Journal of Web Engineering, 001-027.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wang, Q., Yang, L., Kanagal, B., Sanghai, S., Sivakumar, D., Shu, B., ... & Elsas, J. (2020a). Learning to extract attribute value from product via question answering: A multi-task approach. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 47-55).

Wang, Y., Xu, Y. E., Li, X., Dong, X. L., & Gao, J. (2020b). Automatic validation of textual attribute values in e-commerce catalog by learning with limited labeled data. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2533-2541).

Whitehead, K., Zacharia, Z., Prater, E. (2019). Investigating the role of knowledge transfer in supply chain collaboration. The International Journal of Logistics Management, 30(1), 284-302,

Zheng G., Mukherjee S., Dong XL., Li F. (2018). OpenTag: Open Attribute Value Extraction from Product Profiles. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM; 1049-1058. doi:10.1145/3219819.3219839.

Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017) ‘EAST: An Efficient and Accurate Scene Text Detector’. arXiv. Available at: http://arxiv.org/abs/1704.03155 (Accessed: 15 September 2022).Zheng G., Mukherjee S., Dong XL., Li F., OpenTag: Open Attribute Value Extraction from Product Profiles. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM; 2018:1049-1058. doi:10.1145/3219819.3219839.