Production Management and Process Control, Vol. 36, 2022, 116–127
https://doi.org/10.54941/ahfe1001626

Product Data Quality in E-Commerce: Key Success Factors and Challenges

Maciej Niemir and Beata Mrugalska

Poznan University of Technology, Poznan, Poland

ABSTRACT

Digital technologies advanced more rapidly than any other innovation in our life in the last two decades. Such a situation made companies increase theirs affords to offer product data at business-to-business interfaces well-timed and correctly. In order to prevent the risk of poor product data quality, it appeared that they need to identify, analyze and monitor the product data. In this paper, we discuss problems of the development and maintenance of global standards ensuring effective business communication. In detail, we present the exemplary attributes of the products in the registry such as brand name, product image, net content and unit of measure, Global Product Classification, and countries of sale. Their deep analysis allows us to identify challenges and success factors for e-commerce.

Keywords: Basic product attributes, Data quality, E-commerce, Master data, Quality of product data, Product catalog management

INTRODUCTION

Digital technology has changed the face of business enabling e-commerce to flourish and even became a principal point for consumers and businesses in the previous years. This situation also is largely influenced by the pandemic as consumer shifted their behavior to online shopping. According to the Mastercard Economics Institute, customers spent additional $900 billion online in 2020 in comparison to the last two years. Moreover, in the United States, in the same year e-commerce accounted for 14% of total retail sales what shows a double increase in comparison to 2015 (Mathradas, 2021).

Despite the rapid growth of e-commerce in recent years, the successful development and implementation of e-commerce still relies on correct demonstration of products on its platform. In order to be able to offer the products, the corresponding product data has to be provided. In practice, there is a need to arrange various data formats from different sources and remove irrelevant information before showing on the platform. For this aim, data is integrated from multiple sources which require their format unification, schema matching, and information extraction. In spite of the complexity of these tasks, data integration often requires manual adjustments as tools are unsuccessful in automate extract-transform-load data pipelines on non-standard or low-quality data (Schmidts et al., 2020).

In this paper, we present some basic attributes required to provide to product data into e-commerce platforms. For each of them, we made

an analysis to indicate inconsistencies in their understanding. It allowed us to show a lack of commonly available, standardized, consistent data describing products what leads to the development of own solutions for the e-commerce market. However, as the e-market is still young, it seems essential to clearly understand e-product data and provide practice recommendations.

PRODUCT DATA QUALITY

The notion of product data is commonly used to refer to all product-related information (Kropsu-Vehkaperä and Haapasalo, 2011) which can be read, measured, and structured into an appropriate format. It covers physical and functional attributes of the product with its detailed technical information, including also abstract and conceptual information (Sääksvuori and Immonen, 2002). It is vital to emphasize that its quality is a critical issue especially in large databases, but when it is insufficient, it can even have a substantial negative business impact (Byabazaire et al., 2020).

Till now, data quality is still a concept with many definitions as diverse studies theorize data quality by its dimensions (e.g. accuracy, completeness, objectivity, consistency, timeliness, validity, and credibility) based on empirical research, ontological and semiotic framework (e.g. syntactic (structure of data), the semantic (meaning of data), the pragmatic (usage of data) and the social level (shared understanding of the meaning of symbols)) or practitioners' experiences. Generally, these studies lead to the definition of data quality as 'fitness for use' indicating that one data object may vary in different circumstances (Lush et al., 2018). Apart from this conceptualization, data quality metrics can be used to operationalize data quality dimensions and identify the data to be measured (Batini et al., 2015). Other studies refer to procedures and techniques for measuring data quality with interviews and surveys (Price et al., 2008) or validation rules (Fan et al., 2008). In order to identify data defects and formulate data quality metrics in a particular context procedure models and analysis techniques can be also suggested (Heinrich and Klier, 2009). It is proposed that for the design of business-oriented data quality metrics (i.e. metrics for monitoring business-critical data defects) causal relations between data defects, business operations problems, and strategic business goals should be analyzed (Hüner et al., 2009).

The GS1 organization, dealing with the development and maintenance of global standards ensuring effective business communication (commonly associated with the barcode standard on products) in response to the growing problems with the quality of product data available on the Internet, for several years has been obliging all companies using GTIN (Global Trade Item Number), represented as a barcode on the product) for filling the global product registry (known as the Global Registry Platform). By providing widely available verification and data retrieval services (the first one - "Verified by GS1"), the register is intended to help the market to use clear and reliable product data from product manufacturers (GS1 US, 2019). The product in the Registry consists of several attributes:

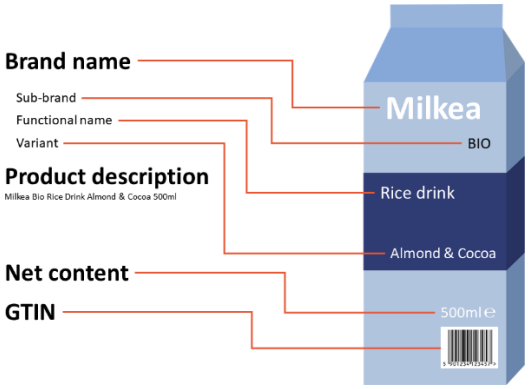

Figure 1: Sample product. Source: Adapted from gs1us.org.

- GTIN (Global Trade Item Number),
- Product description/name that describes the product (repeated by language),
- Brand name (repeated by language),
- Product image (repeated by language),
- Net content and unit of measure (repeated by unit),
- GPC (Global Product Classification),
- Countries of sale / Target market (list).

This Global Product Register will likely be the world's largest register of unique product data. Therefore, it is worth paying attention to the individual attributes of the product listed therein and assessing their uniqueness in the understanding of the current e-commerce market. The article (Niemir and Mrugalska, 2021) mentions two of them: the GTIN number as a key and unique number for the business process support, and the product name/description that uniquely describes the product - important for unambiguous identification by a human. It has been shown that GTIN has no alternatives on a global scale, but unfortunately is not commonly implemented in e-commerce solutions and is not used for unambiguous identification. In the case of the product name - there is no unequivocal interpretation of what it should contain and how it should represent the product. In this article, we will focus on the remaining attributes indicated as primary in the context of the global GS1 register.

METHODS

To achieve the assumed goal, in this paper we selected market representatives, platforms and tools used on the Internet (Table 1). Then, the lists of attribute properties were made, which describe whether the tested product attribute is present, whether it is obligatory, what restrictions it has on entering data, and what standards or best practices their developers refer to. The analysis was made based on:

- analysis of internet traffic statistics in the world in the e-commerce category based on the Similarweb internet traffic ranking (Similarweb, 2021) (statistics from 01/03/2021),

Table 1. E-commerce representatives.

Name	Description
Schema.org	An activity to create, maintain and promote schemas for structured data on the Internet, web pages, in email messages, etc. founded by Google, Microsoft, Yahoo and Yandex, Its vocabularies are established in an open community process by the public-schemaorg@w3.org mailing list and through GitHub. The use of structured data on websites affects search results and the way e-shop product pages are displayed in search engines. It is a widely used standard in e-commerce.
Google Merchant Center	Digital platform where online retailers upload product data that fuels Google Shopping Ads (formerly Product Listing Ads) and provides information about your eCommerce store. Its primary goal is to allow businesses to upload and maintain product information, including pictures and pricing, to be displayed in relevant Google Shopping searches. It is representative data aggregation platforms.
Top marketplaces: Amazon, e-Bay, Allegro	The world's largest online retailer that sells directly or as a marketplace. Number 1 in the e-commerce and shopping Internet traffic ranking in the world (Similarweb, 2021). An online shopping site, the best known for its auctions and consumer-to-consumer sales, but it is also a marketplace, common for online merchants to use as a sales channel. It is in second place in the list of e-commerce and shopping Internet traffic ranking the world (Similarweb, 2021). The most popular shopping platform in Poland and one of the largest e-commerce websites in Europe (Similarweb, 2021).
Price comparision: Ceneo	The most popular Polish price comparison website presenting the offer of over 18,000 online stores. It is in the first place in the Internet traffic ranking in the category of price comparison websites in Poland, and as the second in the world - Similarweb (2021).
Top e-commerce platforms: WooCommerce, Shopify, Magento	Open-source Woocommerce platform built on WordPress is the most often used platform (30%, 35887 websites). Shopify (18%, 22285 websites) is online store builder trusted by over 1,000,000 stores. Magento (9%, 10778 websites) is e-commerce platform built on open source technology that provides online merchants with a flexible shopping cart system. The platforms differ from each other in the business model, the software sharing model (SaaS, self-hosted), the openness of their code, and the possibilities of expanding their functionality through the installation of add-ons.

- analysis of the popularity of using e-Commerce platforms based on BuiltWith (BuiltWith, 2021) data (results published on 04/16/2021),
- own experiences with commonly used e-commerce tools (Google Merchant, Schema.org).

RESULTS AND DISCUSSION

Brand Name

The "brand name" attribute is one of the most important fields that allow you to group and filter products. Although its completion is usually not mandatory, it is commonly used in e-commerce, as evidenced in Table 2. It is also worth noting that the combination of the brand name and the MPN field

Table 2. Comparison of product brand names.

Platform tools	Field type	Max length	Definition
Verified by GS1	Text field	70	The name given by the brand owner which is supposed to be recognizable by the customer. Repeatable by language.
Product - Schema.org Type	Multiple objects		One object or multiple objects. The brand(s) related to a product or service, or the brand(s) maintained by an organization or business person. A brand can have not only a name but also a logo, URL, own ID, and even a motto.
Google Merchant Center	Text field+	70	Necessary for new products, excluding movies, books, and musical recording brands. The field should contain brands recognizable by customers, created by manufacturers. If the product does not have one, it should be the name of the manufacturer or supplier. Cannot contain values: "N/A, Generic", "No brand", or "Does not exist".
Amazon	Text field*	50	A distinctive and recognizable symbol, association, name, or trademark used to distinguish competing products or services. It can refer to a single product, an entire product line, or a company. Used to recognize a vendor's goods or services and to distinguish them from competitors. It must approve the newly introduced brand before it can be used to list products.
eBay	Text field+	65	A exclusive and distinguishable name or symbol used to identify the vendor's goods or services. Brand names can be trademarks and refer to a single product, product line, or even an entire company. The brand name should correctly match the spelling used by the brand manufacturer in the appropriate language. In particular, pay attention to the uppercase and lowercase letters in trademarked brands. Do not include symbols (®, ©, and ™) or abbreviations ("GmbH" and "Ltd.") which are not part of the brand name. Do not use the name of the producer, but the brand name under which the product was specified. Field required: must appear together with the Product: MPN field
Allegro	Dictionary field+		Required field depending on the category (mandatory in most categories). Dictionary field, with the possibility of choosing the option "other" and entering own value.
Ceneo	Dictionary field		Not required. Dictionary field, the values depend on the selected category.
WooCommerce			The field can be freely created, but it is not included by default
Shopify			Not included by default. The field can be added as additional metadata. An existing "vendor" field is often used.
Magento			The field can be freely created, but it is not included by default

** Required field; + required field with exceptions*

reduces the likelihood of confusion in determining the uniqueness of the product, which has been used on the eBay platform. Of course, this still does not provide the same guarantee as using the GTIN number, since no institution oversees and standardizes the MPN number, and there is no obligation to register the brand name in patent offices in all countries, hence the names may repeat, not to mention the problems with typos when entering data.

Table 2 shows that there is no consistency in the maximum permissible field length, while the differences are not large (50-70 characters). The creators of the solutions also agree on the general content of the field (except for the liberal Product approach at Schema.org) - they show that the brand name should come only from the manufacturer, it should be recognizable on the market, it should apply to the product, product line or the entire enterprise. However, it was noted that in the absence of a brand, Google Merchant Center advises to enter the name of the company or supplier – which is a sure way to prevent omitting data in this field but may cause inconsistency in the data, while eBay explicitly forbids entering the name of the manufacturer instead of the brand. Inconsistency and its prevention is the most important lesson that can be drawn from this table. Generally, large multi-product companies use branding systems and product hierarchies where umbrella brands, sub-brands, family lines and product lines allow to distinguish the goods the company sells, while small companies do not need to own a brand at all. As a result, if the data was not entered by the manufacturer, it may not be possible to correctly read the brand name from the product packaging. Consequently, many platforms decided to block the possibility of entering any text values into the "brand name" field in favor of a defined brand dictionary (Ceneo, Amazon, Allegro), and to add additional dictionary elements under strict supervision (Amazon). Although this solution is conducive to maintaining consistency in the platform databases, it is not conducive to the exchange of data between databases, so one of the conclusions, in this case, may be the need to create a central global brand database. Another solution could be to use "Verified by GS1" as a source of brands, however, assuming that GS1 will guarantee the consistency of the database on a global level (currently the owner of the GS1 company prefix in the country where this prefix was registered is responsible for the quality of data) and the field becomes mandatory in the entire platform.

Product Image

The most important parameters of product photos required by e-commerce are listed in Table 3, divided into technical parameters - such as photo size, format, file size, and parameters related to the content - background colors, prohibited content, percentage of coverage in the frame.

Table 3 shows many differences related to the requirement of the image attribute, the permissible number of images, the method of measuring the minimum and maximum resolution as well as the permissible image sizes as well as file formats and their sizes. The TIFF format is noteworthy, as is the less frequently used BMP - presumably appearing in acceptable formats due to the popularity of storing source graphic files in these formats. Note that

Table 3. Comparison of technical aspects of the attribute "Product image".

Platform & tools	Min qty	Max quantity	Min resolution	Max resolution & file size	File format
Verified by GS1	0	Multiple	900px x 900px	4800px x 4800px	JPG, PNG, GIF, TIFF
Product - Schema.org Type	0	Multiple			
Google Merchant Center	1	1+10 (additional)	100px x 100px non-apparel, 250px x 250px apparel	64 megapixels 16MB	GIF, JPG, PNG, BMP, TIFF
Amazon	1	9, only 7 of these will be displayed	500px, min recommended 600px on the longest side	10,000px on the longest side	GIF, JPG, PNG, TIFF
eBay	1	12 for non-motors products 24 for motors products	1,000px (width + height in pixels) – only warning is provided when smallest	15,000px (width + height in pixels)12MB	GIF, JPG, PNG, BMP, TIFF
Allegro	1	15	500px on the longest side	2560px x 2560px2MB	JPG, PNG, BMP
Ceneo	1	Multiple			
WooCommerce	0	Multiple	min recommended 800px x 800px		JPG, PNG
Shopify	0	250	1px x 1px	2048px x 2048px	JPG, PNG
Magento	0	Multiple		1200px x 1200px	JPG, PNG

these formats are not suitable for viewing on the Internet, so this is only a way of passing data sources between databases.

Table 4 indicates whether all tested solutions require similar image content, therefore only platforms that specified it was presented. All platforms specified that the background should be white, plus light gray on eBay and transparent on Verified by GS1. Analyzing the banned materials in the photos - also the compliance was found - no additional logos, promotional texts, watermarks, i.e. anything that would distort the actual image of the product. Unfortunately, the differences begin with the analysis of what is the subject of the photo, in particular the photo - marked as main, first, representative. The product on some platforms may be presented in different perspectives in one image (eBay, Allegro), whereas other platforms prohibit it (Amazon). It can be presented together with the packaging in Allegro, where other platforms do not allow it (Verified by GS1), or allow it conditionally (Amazon). The photo for multipack products in the Google Merchant Center should contain a single product, while eBay only specifies what should be the main element of the photo, in this case, allowing the background. Allegro allows the context of product use, arrangements with the use of the product in certain categories, Amazon has special requirements for displaying clothes and shoes, eBay has a restrictive approach to displaying the human body in many

Table 4. Comparison of the qualitative aspects of the "Product Image" attribute.

Platform tools	Background	Forbidden	Coverage	Multiple views of a single product
Verified by GS1	white	signatures, watermarks	Should be 95%	Not allowed
	High-resolution product image to clearly show the main selling surface of the product. The image should allow retailers to verify the identity of the item. Assumptions are that the primary selling surface is equivalent to the functional front of the product or the Default Front of the item as outlined in the GS1 Measurement Rules (GS1, 2021) (Sect. 4.2) and the GS1 Product Image Specifications (GS1, 2020) (Sect. 1.1).			
Google Merchant Center	white or transparent background	promotional text, watermarks, or borders	75% - 90%	Not allowed
	Show a single unit of the product. If you're using the multipack attribute, the main image should be of a single unit.			
Amazon	white	text, logos, borders, color blocks, watermarks, or other graphics over the top of a product or in the background	Min 85%	Not allowed
	MAIN images must display products outside of their packaging. Boxes, bags, or cases are accepted conditionally only if they are an important product feature. There are special requirements for displaying clothing and shoes.			
eBay	between white to light grey (a light shadow is also acceptable, but mirror reflections are not allowed)	copyright marks, watermarks, reflections, or hot spots, any text that is not part of the original product or box	80%-90%	Allowed, but the two objects in the image must show different angles of the same product.
	A primary image should be a front view of the product, either straight on or at a slight angle. On listings for multipacks, the primary image must clearly show the main product. Images showing body parts are only acceptable if they show body-wear products.			
Allegro	white	logos		Allowed
	You can show a set of products if you sell them together. The item may be displayed next to the original manufacturer's packaging. You can show the product in the context of use, or in an arrangement in selected categories.			
Ceneo	white		should fill the entire space of the photo and be in its center	

Table 5. Comparison of product weight attribute.

Platform tools	Content	UoM
Verified by GS1	Net	UN/CEFACT Common Code
Product - Schema.org Type	Gross	UN/CEFACT Common Code
Google Merchant Center		
Amazon	Gross	[gr], [kg], [lb], [mg], [oz]
eBay		
Allegro	Gross	[kg]
Ceneo	Gross	[kg]
WooCommerce		
Shopify	Gross	[g], [kg], [oz], [lb]
Magento		

categories. In addition, the platforms distinguish the percentage of product coverage of the photo at a different level, unfortunately, these levels do not match.

Net Content

"Verified by GS1" defines the net content field as the quantity of product which is in the package with the unit of measure, usually printed on a label to sale on the market. This field is optional, repeatable by the unit of measurement (UoM). UoM code list is based on Recommendation 20 of the United Nations Economic Commission for Europe (UNECE). None of the compared e-commerce solutions implemented a "net content" field, only an optional gross weight (Amazon, Allegro, Ceneo, Shopify also a Product property at schema.org), presumably mainly to prepare the shipping process. In some solutions, it was also necessary to enter the details of the packaging itself. As for the implementation of the gross weight field, there were also discrepancies in the scope of the possibility of using units of measurement. Table 5 shows the results of the comparison.

The tested e-commerce solutions did not require loading fields related to the net content of the product. The main data needs focused on package dimensions and weight, and product weight (gross). For this reason, a separate column "gross weight" was included in Table 6.

Product Classification

The above examples show that e-commerce solutions are usually not based on global standards. Categories are created freely by users or must be adapted to the requirements of the platforms (except for Google Merchant Center, where AI defines the category based on the categories created by the user). These, after assigning a classification code, in some cases force the entering of additional fields - specifying the user's offer, or due to legal requirements (e.g. restrictions related to the sale of alcohol, food, or clothing). It is worth mentioning that there are many competing standards in the world, for example:

Table 6. Comparison of product classifications.

Platform tools	Taxonomy	User-defined classification description
Verified by GS1	Global Product Classification (GPC)*	
Product-Schema.org Type		Any category for the item as a string, URL, or another object. In the case of text selection, greater signs or slashes indicate a category hierarchy
Google Merchant Center	Google product category	Product type with full category defined by the user. For example, include "Home Women Dresses Maxi Dresses" instead of "Dresses"
Amazon	Amazon product category*	
eBay	eBay Category*, eBay Category (2)	Store Category, Store Category 2 – own defined
Allegro	Allegro classification*	
Ceneo	Ceneo category*	
WooCommerce		Own defined category
Shopify		Own defined product type
Magento		Own defined category

** Required field*

GS1 GPC (Global Product Classification), UNSPSC (United Nations Standard Products and Services Code), ECLASS, but - as shown in Table 6, they are not widely used in e-commerce.

Other Attributes

None of the analyzed e-commerce solutions had explicitly defined fields "country of sale" and "target market" at the product management level. This is because e-stores generally define the countries of sale or the sales restrictions at the shipping level, while the language is usually the language of the e-platform.

The attributes described and compared in this publication relate directly to the product, and not to the offer, or logistics data. For this reason, fields such as packaging dimensions, detailed offer description, or offer price (not to be confused with the manufacturer's suggested price) were not compared. However, it should be noted that these fields, as well as specific fields, depending on the type of product, are required or at least desirable to be completed by selected platforms.

CONCLUSION

The presented research results show that the attributes, indicated in the document as basic, are important, while the need to enter them, as well as the detailed requirements of e-platform operators - differ from each other. The differences are significant, they rely on the use of different standards of dictionaries that cannot be easily mapped, different, sometimes contradictory

requirements for product images, as well as descriptive texts, the content and meaning of which are also not the same. As a result, it is currently not possible to use one set of product data in such a way as to meet the requirements of all e-platforms, which translates into the need to adjust data each time in the event of extending the scope of sales. This, in turn, means that it is often not done by product manufacturers themselves, but by sellers in the supply chain, which may result in errors and reduced sales. The solution to this problem is the full standardization of all product attributes and maintaining data without the possibility of creating changes in the data by subsequent links in the supply chain, but such standardization may take many years, and it may simply not be feasible.

ACKNOWLEDGMENT

This research was funded by Poznan University of Technology, grant number 0811/SBAD/1053.

REFERENCES

Allegro (2021) Wystawianie i edycja oferty. https://allegro.pl/pomoc/dla-sprzedajacych/wystawianie-i-edycja-oferty.

Amazon (2021) Amazon seller central. https://sellercentral. amazon.com/.

Batini, C., Rula, A., Scannapieco M., Viscusi, G. (2015) From Data Quality to Big Data Quality, Journal of Database Management 26(1):60-82.

Byabazaire, J., O'Hare, G., Delaney, D. (2020) Data quality and trust: Review of challenges and opportunities for data sharing in iot. Electronics, 9(12), 2083.

BuiltWith (2021) Distribution for websites using e-commerce technologies. https://trends.builtwith.com/shop.

Ceneo (2021) Ceneo – Instrukcja tworzenia pliku XML. https://www.ceneo.pl/poradniki/Instrukcja-tworzenia-pliku-XML.

eBay (2018) Catalog best practices guide. https://developer.ebay.com/devzone/merchant-products/catalog-best-practices/content/index.html.

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

Google (2021) Product data specification - google merchant center help. https://support.google.com/merchants/answer/7052112.

GS1 (2020) Gs1 product image specification standard. https://www.gs1.org/docs/gdsn/Product_Image_Specification.pdf.

GS1 (2021) Gs1 package measurement rules standard. https://www.gs1.org/docs/gdsn/3.1/GS1_Package_Measurement_Rules.pdf.

GS1 US (2019) Verified by GS1 frequently asked questions. https://www.gs1us.org/industries/emerging-topics/verified-by-gs1.

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

Hüner, K., Schierning, A., Otto, B. (2011) Product data quality in supply chains: the case of Beiersdorf. Electron Markets 21, 141–154.

Kropsu-Vehkaperä H., Haapasalo H. (2011) Defining product data views for different stakeholders". Journal of Computer Information Systems.

Lush, V., Lumsden, J., Bastin, L. (2018) "Visualisation of trust and quality information for geospatial dataset selection and use: Drawing trust presentation comparisons with B2C e-Commerce" in: IFIP International Conference on Trust Management (pp. 75–90). Springer: Cham.

Magento (2020) Magento user guide. https://docs.magento.com/user-guide/ catalog/product-create.html.

Mathradas A. (May 24, 2021) Three traits of a successful e-commerce business, Forbes: https://www.forbes.com/sites/forbesbusinesscouncil/ 2021/05/24/three-traits-of-a-successful-e-commerce-business/

Niemir M., Mrugalska B. (2021) Basic product data in e-commerce: specifications and problems of data exchange, European Research Studies Journal 24(5), 317–329.

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

Schmidts O., Kraft B., Winkens M., Zündorf A. (2020) "Catalog integration of low-quality product data by attribute label ranking", in: Proceedings of the 9th International Conference on Data Science, Technology and Applications, pp. 90–101.

Similarweb (2021) Top sites ranking for e-commerce and shopping in the world. https://www.similarweb.com/top-websites/category/e-commerce-and-shopping/.

Sääksvuori A., Immonen A. (2002) Tuotetiedonhallinta – PDM. Talentum Media Oy

W3C Schema.org Community Group (2021) https://schema. org/Product.

WooCommerce (2021) Adding and managing products - woocommerce docs. https://docs.woocommerce.com/document/managing-products/.