ALTERNATIVE SOURCES OF INFORMATION TO SUPPORT A SET OF STATISTICAL INDICATORS OF THE BOOK PUBLISHING

Olena Horobets,

Postgraduate Student, National Academy of Statistics, Accounting and Audit, Ukraine

ABSTRACT

Issues of supporting the book publishing as an economic activity by a set of statistical indicators are investigated. It is found out that the existing set of statistical indicators does not meet the needs of researchers and practitioners, which is the case of not only Ukraine, but the global book publishing area. The case of the Ukrainian book publishing is taken for analysis to identify core problems faced by this industry. It is emphasized that a comprehensive study of the book publishing industry and presentation of the statistical information with high level of quality and aggregation requires the involvement of new alternative sources of data, of which big data should be highlighted. The component of scientific novelty is that an updated system of statistical indicators is proposed for the first time, with eight modules of sources of statistical information as alternative ones: questionnaires, electronic books, digital libraries, websites of publishers and bookstores, electronic diaries of reading (Goodreads as an example), social networks (Instagram, Facebook, Telegram (open channels)), video hosts (YouTube being the most popular one), and blogs. It is stressed that all the modules of alternative data must be involved for obtaining reliable data, where output data will be processed anew and have direct and reverse links, which will require the use of neural networks with efferent type of links. This statistical support to the book publishing industry is an innovation designed to meet urgent needs of the public and official statistics.

Introduction. One of the key terms imposed on the statistical reporting and to the information contained therein is the quality that embodies a diverse set of criteria, e.g. completeness, accuracy, reliability, consistency of data, and their timely processing. The availability of high-quality statistical reporting enables for carrying out a comprehensive assessment of the industry's performance, finding shortcomings in the operation, monitoring the dynamics of industry indicators and prediction of its future trends. Once the reporting is imperfect, there may be mistaken decisions resulting from the missing data that may signal future (or current) problems in the industry.

A form of a statistical observation is the statistical reporting that involves filling specially approved documents (reports) by which official bodies receive information from respondents required to assess their performance [1].

Eurostat draws attention to the two main problems in statistical collection of information about the sector of culture:

1. The diversity of statistical organizations charged with culture statistics in Europe. This means that the performance of culture statistics significantly varies across countries, with the responsibility for the organization of culture statistics shared by institutions located at various levels of administration

(cultural centers, administrations, ministries of culture, statistical institutes), which practice various ways of statistical data collection and have distinct views of the responsibility for its quality.

2. Data production is country-specific in terms of collection methods, periodicity, coverage and sources [2].

As regards Ukraine, the following shortcomings in the existing statistical reporting of the book publishing can be highlighted:

1. The Laws of Ukraine "Book Publishing Activities" and "The Obligatory Copy of Documents" are violated by the publishers themselves, who do not send an obligatory copy in a timely manner. As a consequence, the Public Institution "Book Chamber of Ukraine named after I. Fedorov" and other legally fixed recipients do not receive 5 to 10% of obligatory copies from the publishers. Meanwhile, by UNESCO criteria, missing data on more than 2% of the copies is regarded as a loss in the national treasure [3].

2. Neither of the statistical forms requires reporting of digital formats of editions (i. e. electronic books or audio books) as an integral part of book publishing, and general recommendations on electronic books, given only in ISO 9707, have not been implemented so far in the domestic reporting.

3. A content review of "Report on Sales and Stocks of Goods in the Selling Network" reveals that the report does not distinguish between books and periodicals, meaning that the data on bookselling are not reported separately. It follows that the bookselling and selling of periodicals cannot be analyzed separately.

4. The book publishing today is assessed by a too narrow range of kind indicators (i. e. only two) that are not sufficient for an extensive description of this industry.

Results. The above mentioned calls for constructing a set of statistical indicators of the book publishing in a way to make it topical and capable of providing a multifaceted portray of the industry (Figure 1[1]).

The first module of indicators (Primary – Absolute – Quantitative) (Figure 1), for which there are no data on the number of electronic editions and the respective number of downloads, which, if available, would give a grasp of the demand for an electronic book, but absence of these data leaves no chance for comparisons or forecasting of future trends in the book of such format.

The data in the second module (Primary – Absolute – Qualitative) enable for the assessment of the qualitative content of the book market. ISBN allows for the assessment of qualitative parameters of an edition, but given the unfair treatment of the registration requirements by publishers this system cannot provide the comprehensive information now. When reporting, publishers include details on topic and purpose of the edition, publisher details, language and territorial details. Like in the first module, in the second one no details on the edition's format (printed or electronic) or the edition's cover (hard cover, paperback, semi-paperback) can be found, although they enable one to reveal readers' preferences regarding the edition's presentation. As far as the fiction as a separate type of editions is concerned, reports do not inform the genre of a fiction work (detective, thriller, novel, erotic work (marked 18+) etc.), which, if provided, could give both publishers and researchers a better grasp of readers' likes. Now the publishers can only rely on their own observations or questionings (conducted in bookstores, online platforms or by means of e-mail).

The third module of this set of statistical indictors (Secondary – Relative – Quantitative) contains the indicators most of which have not existed so far. But recent studies (conducted by the Sociology Institute of the National Academy of Sciences of Ukraine on line of the project "Ukrainian Reading and Publishing Data 2018") revealed a piece of information on the population's reading. But because these studies were conducted ad hoc by various institutions, it should not be expected that they would be followed in future. Also, there is not data on average circulation, average book price (including price variation by type or genre of edition), break down of editions by genre and shares of editions by type or purpose in the total number of editions.

The fourth module (Secondary – Relative – Qualitative) contains inexistent indicators describing an edition. Details on the approximate duration of reading, its periodicity and reading habits can be obtained by means of questioning, but they will not be reasonably accurate due to subjective reasons. At the same time, an electronic book allows for obtaining much more reliable information.

---

[1]On Figure 1 the regular font is used to signify the industry indicators that are produced now, whereas the indicators that are missing in the statistical reporting right now are shown in italics.

Thus, data on duration of reading or its periodicity (daily periodicity in particular) can be feasibly collected by use of algorithms and given the readers' agreement on processing of cookie files of an electronic edition; this can be possibly done due to the activated status of an electronic book, i. e. the time when this electronic book continues to be activated in a specific publisher and in a particular time span. This also applies for reader habits, such as beginning of the reading from the last page, "the reader's rule of the first fifty pages", number of finished, unfinished or re-read books. Besides that, it is possible to obtain a response on the question whether or not people really read what they buy or they put a book on electronic shelves without opening it.
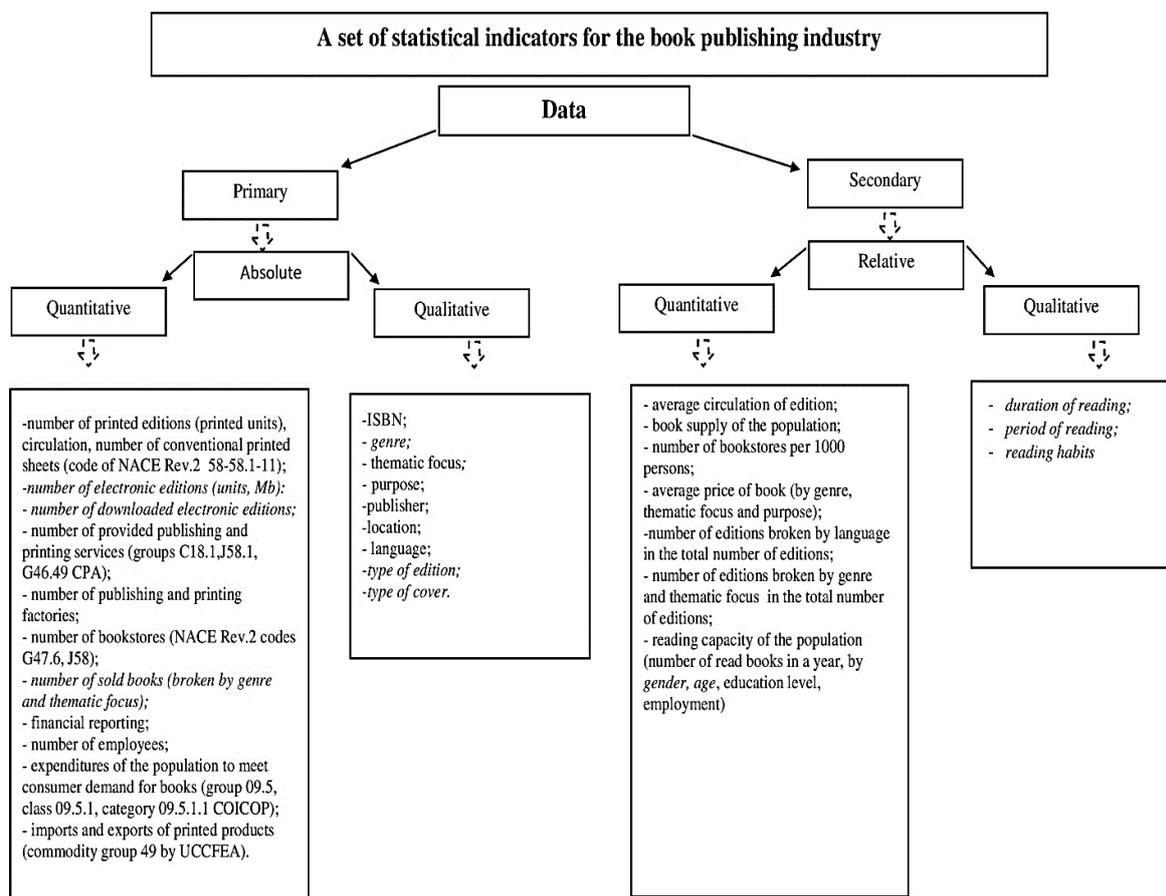
Fig. 1. A set of statistical indicators.

Notes: NACE Rev.2 – Statististical classification of economic activities in the Europeenne Communaute/ Nomenclature statistique des Activites economiques dans la Communaute Europeenne; UCCFEA – Ukrainian Classification of Commodities for Foreign Economic Activities; CPA– Statistical Classification of Products by Activities in the European Economic Community; COICOP – Classification of Individual Consumption by Purpose).

Source: author's development.

So, a review of the statistical reporting of the book publishing industry allows for the conclusion that its shortcomings revealed in course of the analysis have adverse consequences such as the following: spreading of "shadow printing" and provision of unreliable information on numbers of editions and their circulation; wide-scale practices of piracy and insecurity of copyrights; inadequate portraying of the book ecosystem.

While there are only few publications devoted to a review of book publishing activities and even fewer publications about their economic aspects, studies dealing with the implementation of big data in the book publishing statistics are unlikely to be found at all. In the fifth fundamental principle of the official statistics it is stated that data for statistical purposes can be collected from any sources on the basis of both statistical observations and administrative reports, and statistical departments must choose a source with consideration to quality, timeliness, costs and respondent burden [4].

The abovementioned makes us proceed with a review of the role and significance of big data in the official statistics and their applications in other kinds of activities, so as to come out with approaches to implementing big data in the book publishing statistics, in order to improve it.

The topic of bid data was included in the agenda of the official statistics community in 2013–2014. Almost at the same time three new initiatives were launched at national, regional and global level: NIST, UNECE, and UN Global Working Group on Big Data for Official Statistics. O. Osaulenko points out that "the well-known sources of big data can by classified by the following groups: administrative data (electronic medical cards, banking details), transactions or business information (transactions by credit and deposit cards), data from sensor catchers (data from satellite surveys, road radars), data from mobile sensor devices (information from mobile phones GPS), behavioral data (Internet inquiries), information about individual and social opinions (comments in social networks)" [5].

At the Conference of European Statisticians (2019), a seminar "New data sources – accessibility and use" was held, where additional groups of big data were proposed, namely

1. Data of mobile telephony;
2. Websites of electronic commerce and business websites;
3. Crowdsourcing (Open Street Map, Instagram, etc.) [6].

In spite of the above, the issue of implementing big data as an alternative source of statistical information in the official statistics production still remains open. In view of this, it is important to realize that big data is a chaotic flow of data, containing information on a variety of walks of human life in form of multiple disordered structures of data. This conglomeration of data has already been labeled as "flood" that requires special technologies (software packages first and foremost) and highly skilled specialists for their effective processing within an allowable time span [7].

O. Korepanov, when investigating the issue of "smart city" development in Ukraine, clarified that "such data involve a logical scheme that allows for obtaining information for analysis" [8].

V. Sarioglo points out that the increasing interest to issues of implementing "big data" in the statistics is related, by far and large, with the significant commercial success of this approach in the U.S. [9].

O. Osaulenko emphasizes that the project of big data integration can be implemented with reliance upon solutions of the set of fundamental problems:

- On what particular kind of big data the official statistics must focus, bearing in mind its social assignments and functions;
- What is the way by which the official statistics can use big data and what it needs to do this;
- What is the way by which big data are able to help perform more accurate and timely statistical assessment of social, economic and environmental phenomena [5].

The telecommunication provider "Kyivstar" operating in Ukraine makes analyses of data on 26 million subscribers, use of 10 Mb of traffic per day and 90 million of calls per day, to construct the sets of characteristics by three groups: device and traffic details, financial details, mobility details. These characteristics are employed in creating client portraits, look-alike model, target communication, heatmaps and location-based analytics, scoring. In parallel with this, services to big and small business are provided, including developments of trigger marketing technologies, developments of cloud infrastructure for storage, analysis and visualization of data, etc.

A successful story of electronic commerce using big data in Ukraine is Internet stores like Rozetka, Makeup, Yakaboo, etc. Although the three stores differ by assortment of goods and purpose (Rozenka is focused on selling technical devices; Makeup – on perfumery and cosmetics; Yakaboo – on books and stationary), they all operate by one principle, which is monitoring of the customer focus, because when you are a registered customer and you visit one of the abovementioned websites, you will immediately receive a e-mail offer to buy a commodity you have just looked at.

So, a review of practices used by giants of electronic commerce suggests that their good business stories can be adapted by any industry.

The book publishing sector in Ukraine also has its story of successful use of big data in order to conduct a readers' activity review. It was earlier in 2020 that on the commission of the Institute of Book, the Center of Content Analysis "selected public posts in social networks (Facebook, Instagram, Twitter), video in YouTube and messages in public channels of Telegram messenger for the period from January 1 till July 31, 2020, which contained the word "book(s)" or word combinations of genre names with the verbs "read", "publish" etc. or names of key book-related events (Bookforum, "Book arsenal" etc.) and had at least one link (like, dissemination, comment, re-twit etc.) or at least 10

lookups, and were written by users who signified Ukraine as their country of residence or wrote a post in Ukrainian. In this way 1 387 163 content units were identified, of which 3 000 units of content were selected for coding by use of a special sample. In the process of coding 2212 relevant mentions of reading were identified, with the sample error being +/-2% given the confidence level of 95%".

Three units of coding were used in processing the obtained data: messages in social networks and messengers (search of book titles, authors, formats etc. was carried out); titles of books mentioned in messages (by identifying book author, genre etc.); profiles of message authors (their data in the section "about me", and the content of public messages throughout one month since the day of coding)" [10].

The abovementioned practice of using big data gives another evidence of the topicality and efficiency of such studies. It should be noted that processing of big data, like any other data set, requires a well adjusted mechanism with clear division of tasks and good awareness of the end goal.

Processing of big data on the book publishing can be referred to as the efferent type of links in an artificial neural network (or a network with direct and reverse links) (Figure 2), when to the entry of neuron of i layer its input signal is given, i. e. this neuron strengthens or weakens the signal, transformed by its activating function, by which its marginal activation status is achieved [11].

It should be noted that factor analysis, correlation analysis, ranking, etc. can be used at entries or exits of neurons in case of need, in order to achieve deeper detailing of conclusions.
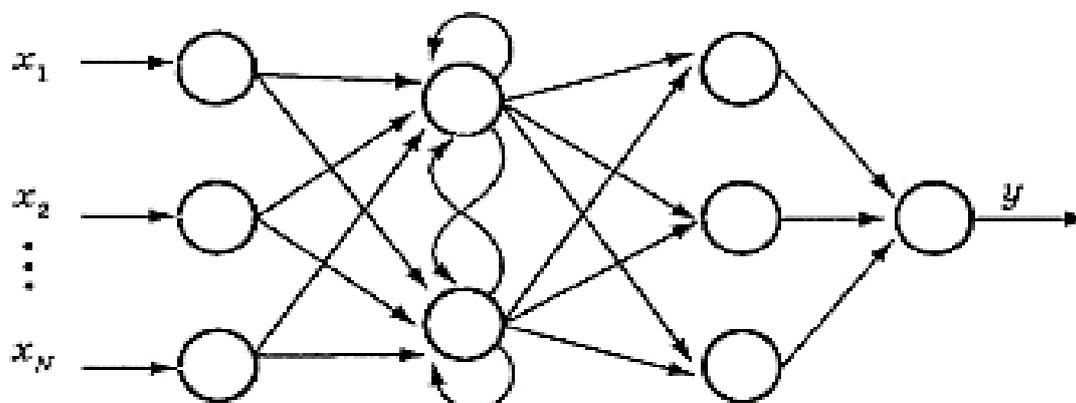
Fig. 2. An artificial neural network with direct and reverse links
Source: [11]

The abovementioned type of a neural network is regarded as the dynamic one, because the reverse links (loops) make the neuron entries modify in time, thus changing the network status [11].

Such processing of data has strong and weak sides. The weak side is that a great many dark data occur in this process. The strong side is that the more intersection points (neural network nodes) occur in the process of data collection, the better is its impact on the quality of output information, and the ability of a neural network to identify the links between seemingly varying and unrelated parameters enables for a concise depiction of big data, if the actual correlation between data is high. A large conglomeration of dark data can be avoided given the clearly defined end goal of data collection, which must determine the construction of data modules and the logic setup in the training of neutral network with correctly fixed input vectors according to the weight and procedure.

However, as put by A. Kononiuk, "use of neural networks requires a number of conditions to be met by the designer: a set of data providing for the problem description; understanding of the essence of the addressed problem; choice of adder function, transfer function and methods of training; understanding of instrumental devices by the designer; the appropriate capacity of processing and the designer's skills beyond the boundaries of traditional computations" [11].

Given the obvious advantages of big data, their weaknesses need to be remembered: financing, methodological support, creating new policy for data management, supply of statistical services with technical devices and software packages, inaccuracy of algorithms, unreliability of computerized processing, risk of data loss, flaws in the regulatory aspect of big data issue, violation of the confidentiality of private data, etc.

When analyzing the issue related with confidentiality of private data as a subcategory of big data, it should be noted that OECD defines the following categories of personal data: content created by user (such as blogs, photos, comments, video); data on activities or behavior in Internet

(downloaded references, information about the shopping basket in e-shops); social data (contacts on websites containing social networks, dating websites); local data (geographic location, IP address); demographic data (age, gender, sexual preferences); identification data of the official nature (bank account, number of social insurance) [12].

Article 8 of the Charter of Fundamental Rights of the European Union says that such data need to be processed fairly and for defined goals given the stakeholder's agreement or on other legal basis fixed by the law. Each one has the right to access the data collected about him/her, and the right for their correction [13]. Principle 6 of the Fundamental Principles of Official Statistics specifies that the personal data collected by statistical departments for preparation of statistical information, irrespective of whether they are concerned with physical persons or legal entities, must have strictly confidential nature and be used only for statistical purposes [4].

Official statistics bodies must improve data collection technologies, ensure storage, protection and confidentiality of statistical information as part of their professional effort, with primary data used for statistical purposes in aggregated, depersonalized form guaranteeing their confidentiality.

So, eight modules of alternative sources of statistical data should be proposed for information support of the book publishing: questionnaires, electronic books, digital libraries, websites of publishers and bookstores, electronic diaries of reading (such as Goodreads), social networks (Instagram, Facebook, Telegram (open channels)), video hosts (most popular one being YouTube), blogs (Table 1).

Table 1. Modules of alternative data for the assessment of book publishing

Questionnaires	Electronic books	Digital libraries	Website of publisher or bookstores	Electronic diaries of reading	Social networks	Video hosts	Blogs
A	B	C	D	E	F	G	H
Depersonalized data							
Buyer's details: -e-mail; -phone number; -age; -family status; -place of residence; -hobby; -existence of pets and children; -favorite genre; -publishers which books prevail in the home library	*Quality of reading:* -number of downloaded books (periodicity of downloading); -number of read pages; -periodicity of reading; -duration of reading (e. g. average duration in periods); -genre; -author; -process of book reading; -habits of reading	*Reader preferences:* -genre; -author; -number of downloads; -number of abstract lookups; -number of lookups of the introductory material to a book; -popularity of author; -popularity of book; -library router; -time spent in the library (e. g. average time in a period); -period of the readers' activity	*Baskets and bookstores:* -preferred genre; -website router; -saved books; -purchased books; -average basket bill; -chosen books; -purchased books; -competitiveness of bookstore; -income of bookstore; -popularity of author; -popularity of book; -periodicity of order; -period of the readers' activity	*About books:* -review on read book, book author, publisher; -book rating; -popularity of book and author; -recommendations	*About books, authors, publishers:* -popularity of author, book, publisher (number of mentions)	*About books, authors, publishers:* -popularity of author, book, publisher (number of mentions); -existence of a feature film or a serial based on a book	*About books, authors, publishers:* popularity of author, book, publisher (number of mentions)

Source: author's development.

Work with big data should be started with scrutinizing best practices of the giants of electronic commerce, in order to avoid their mistakes and benefit from previous achievements to the maximally possible extent.

Conclusions. This study of the statistical reporting as part of information support of the book publishing revealed the main industry-specific problems: wide-scale practice of "shadow printing", widespread piracy and inadequate portraying of the book ecosystem. It is found that the existing statistical support of the book publishing is imperfect and built on a narrow range of kind indicators that are not capable to give its comprehensive description at the current phase of development. Based on this analysis, an up-dated set of statistical indicators for the book publishing is proposed, which implementation requires the involvement of alternative sources of statistical data and innovative approaches to their analysis.

Results of the study give ground for the suggestion that the statistical assessment of the book publishing in Ukraine can be improved first and foremost through identifying new data sources. Turning back to fundamental issues of big data integration, it can be concluded that they could be dealt with in the context of book publishing statistics.

REFERENCES

1. Osaulenko O. H. (Eds.) (2012). *Statystychnyi slovnyk [Statistical glossary]*. Kyiv: "Information-analytical agency", p. 138 [in Ukrainian].
2. European Commission. (2012). ESSnet-Culture Final Report, Luxembourg. Retrieved from: https://ec.europa.eu/assets/eac/culture/library/reports/ess-net-report_en.pdf.
3. Derzhavnyi komitet telebachennia ta radiomovlennia. Departament informatsiinoi polityky. Nedostavliannia oboviazkovoho prymirnyka vydan zahrozhuie vtratoiu natsionalnoho nadbannia [The State Committee on Television and Boradcasting. Department of Information Policy. Not providing the obligatory copy of editions threatens with loss of the national treasure]. 28.04.2010. Retrieved from http://comin.kmu.gov.ua/control/uk/publish/printable_article?art_id=74701. [in Ukrainian].
4. United Nations Statistics Division. Fundamental Principles of Official Statistics (A/RES/68/261 from 29 January 2014). Retrieved from: https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.
5. Osaulenko O. H. (2017). *Ofitsiina statystyka v systemi natsionalnoi informatsiinoi bezpeky [The official statistics in the system of the national information security]*. Kyiv: "August Trade", pp. 259–260 [in Ukrainian].
6. United Nations. Economic Commission for Europe. (26–28 June 2019). New data sources for official statistics – access, use and new skills. *Note by United Nations Statistics Division*. Paris. Retrieved from: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2019/ECE_CES_2019_41.pdf.
7. Eastern Partnership Civil Society Facility. (November, 2018). The Data Deluge: Blessing or Curse? *Ideas from DataFest Tbilisi*. Retrieved from: http://eapcivilsociety.eu/news/project-news/the-data-deluge-blessing-or-curse-ideas-from-datafest-tbilisi-november-2018.html.
8. Korepanov O. S. (2018). *Metodolohichni zasady statystychnoho zabezpechennia upravlinnia rozvytkom «rozumnykh» mist v Ukraini [Methodological foundations for statistical support to management of "smart cities" development in Ukraine]*. Kyiv: "Information-analytical agency", p. 228 [in Ukrainian].
9. Sariohlo V. H. (2016). "Velyki dani" yak dzherelo informatsii ta instrumentarii dlia ofitsiinoi statystyky: potentsial, problemy, perspektyvy [Big data as a source of information and a tool for the official statistics: potentials, problems, prospects]. *Statystyka Ukrainy – Statistics of Ukraine*, 4, 12–19. Retrieved from http://nbuv.gov.ua/UJRN/su_2016_4_5. [in Ukrainian].
10. Ukrainskyi instytut knyhy. Chytannia v konteksti mediaspozhyvannia: vplyv karantynu na chytatsku povedinku ukraintsiv [The Ukrainian Institute of Book. Reading in the context of media consumption: the impact of the quarantine on the reading behavior of Ukrainians]. Retrieved from https://drive.google.com/file/d/15eZw4iK-ldrXewQLFQ63el_8nBE-GFqU/view. [in Ukrainian].
11. Kotsovskyi V. M. (2013). *Neironni systemy: konspekt lektsii [Neural networks: lecture notes]*. Uzhhorod, p.13. Retrieved from https://dspace.uzhnu.edu.ua/jspui/bitstream/lib/16450/1/%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%BD%D0%BD%D1%96%20%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B8.%20%D0%9B%D0%B5%D0%BA%D1%86%D1%96%D1%97.pdf. [in Ukrainian].
12. OECD. (2013). Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value, OECD Digital Economy Papers No. 220, OECD Publishing, Paris. Retrieved from http://dx.doi.org/10.1787/5k486qtxldmq-en.
13. European Union Agency for Fundamental Rights. EU Charter of Fundamental Rights. Article 8, Protection of personal data. Retrieved from https://fra.europa.eu/en/charterpedia/article/8-protection-personal-data.