ОБЗОР НА ВИДОВЕТЕ ИЗТОЧНИЦИ НА ДАННИ

Мартин Дойнов[1]

[1] Икономически университет – Варна/Катедра „Информатика“, Варна, България
martin.doynov@ue-varna.bg

РЕЗЮМЕ

Събирането на информация и извличането на знания от нея е процес, залегнал в основата на човешкото развитие. С еволюцията на технологиите, методите за боравене с данни непрекъснато се надграждат и усложняват. Въпреки че в дигиталната епоха разполагаме с всякакви нови инструменти, подпомагащи процеса на подбор на данни, както и практически неограничени количества информация, с големите обеми нарастват както трудността да се намери съществено, така и рискът да станем жертва на неверни записи. Ако нямаме знанията как да филтрираме данните правилно и да извлечем добавена стойност от тях, те няма да ни служат. Именно това налага познаването в дълбочина на източниците на данни и добрите практики, които всеки изследовател трябва да познава, за да подбира само най-подходящото за своите проучвания. Тази изследователска работа се провежда като част от проект № ПНИ-КС24-04-DCAAITM.

КЛЮЧОВИ ДУМИ: *данни, източници, видове, информация, извличане*

AN OVERVIEW OF THE DATA SOURCE TYPES

Martin Doynov[1]

[1] University of Economics – Varna/Department of Informatics, Varna, Bulgaria
martin.doynov@ue-varna.bg

ABSTRACT

Collecting information and extracting knowledge from it is a cornerstone process of humanity's development. Along with the evolving technologies, the methods for handling data are also built upon and further sophisticated. Despite the variety of modern tools, helping researchers in the process of data collection, as well as the practically unlimited quantity of information at hand, the difficulty of finding substantial information and the risk of falling victim to faulty records increases due to the sheer volumes of available knowledge. If we do not possess the skills to filter data correctly and extract added value from the conclusions, they will be of no use. This is what necessitates the need for deepened understanding of data sources and the best practices essential to any researcher in order to pick only the most suitable material for their research. This academic work was conducted as part of project № ПНИ-КС24-04-DCAAITM.

KEYWORDS: *data, sources, types, information, extraction*

ВЪВЕДЕНИЕ

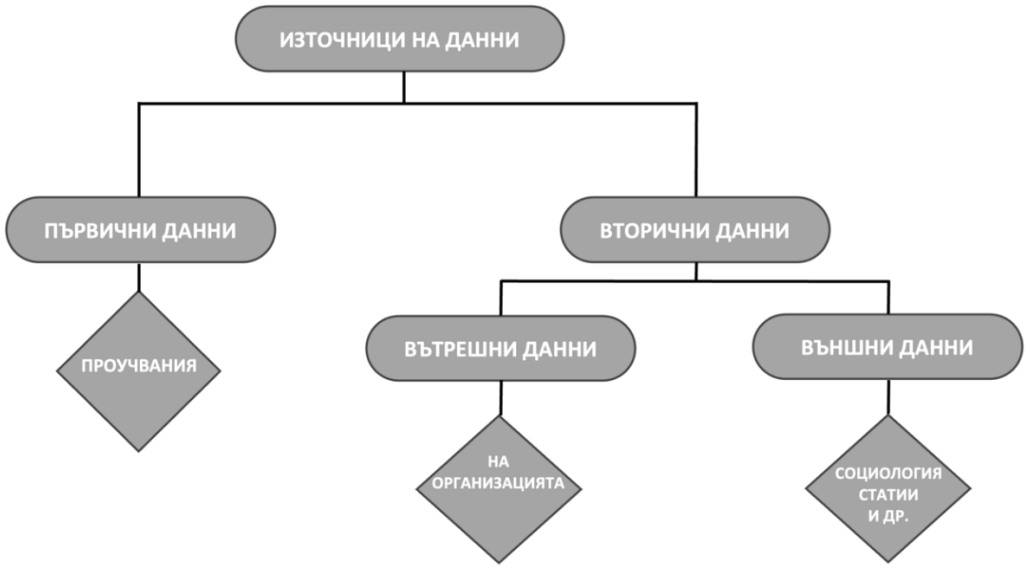
Преди предприемане на стъпки към обработката на данни, трябва да се състави план за техния подбор и да се оценят източниците, които ще бъдат използвани. Не винаги е нужно да разполагаме с големи количества данни, но и тяхното съдържание и структура да бъдат подходящи за нашите цели. Ако данните съдържат грешки, или тематиката им не съвпада с тази на нашето проучване, целият процес се превръща в безпредметен. Затова е нужно да познаваме различните типове източници и да знаем какво е нужно, за да спазваме правилна хигиена на данните.

Цел на доклада е да запознае аудиторията с различните видове източници на данни и да ги съпостави според признаците, които ги отличават един от друг. Ще бъдат изброени методи за събиране на данни, заедно с примери за приложението им в различен контекст. Основен разгледан източник на данни е световната мрежа Интернет, поради своята достъпност, популярност и непрекъснато увеличаващ се обем от данни. Особено внимание се обръща върху съвременните проблеми при подбора на информация, свързани с качеството на данните и обработката им, както и етичните и моралните проблеми, съпътстващи употребата на лични данни.

1. ТИПОВЕ ИЗТОЧНИЦИ НА ДАННИ

В повечето изследвания източниците на данни се делят на два основни типа – **ПЪРВИЧНИ** и **ВТОРИЧНИ** (вж. фиг. 1). Категоризират се според начина им на събиране. Както подсказва името, първичните данни са оригинални и се събират за първи път, докато вторичните данни вече са събрани или създадени от други. Основното отличаващо е, че първичните данни са факти и са оригинално създадени, а вторичните представляват анализ и интерпретация на първични данни.

Първичните данни се събират с цел решаване на конкретен проблем и отразяват реалността в момента. За разлика от тях, вторичните данни често се отнасят до минали събития и са събрани за цели, различни от разглеждания проблем. Терминът „първични данни“ се отнася до информация, създадена от изследователя, докато вторичните представляват вече налични данни, събрани от организации или агенции (Mesly, 2015). Процесът на събиране на първични данни изисква значителни усилия и време. За разлика от това, събирането на вторични данни е значително по-бързо и лесно. Източници на първични данни могат да бъдат проучвания, експерименти, наблюдения, въпросници и лични интервюта. В същото време, вторичните данни обикновено се черпят от източници като правителствени публикации, уебсайтове, книги, статии в списания и вътрешна документация (Ajayi, 2023).



Фигура 1. Разделение на видовете източници на данни

Източник: geeksforgeeks.org (2024)

Основните разлики между първичните и вторичните данни са изведени в таблица 1.

Таблица 1.

Сравнение на първичните и вторичните данни

Признак за сравнение	Първични данни	Вторични данни
Определение	Данни, събрани собственоръчно от изследователя	Данни, събрани предварително от някой друг
Актуалност	Събрани непосредствено преди проучването	Събрани в неопределен период в миналото
Сложност на събиране	Изисква голяма ангажираност от страна на проучвателя	Опростено
Разход	Скъпо	Икономично
Време на събиране	Дълго	Кратко
Специфичност	Винаги специално подбрани за целите на проучването	Може да не са подбрани специално за целите на проучването
Формат	Необработени	Обработени
Точност и надеждност	Висока	По-ниска

Източник: (Wagh, 2024). Адаптирано от автора

ПЪРВИЧНИТЕ ДАННИ произлизат от първични източници на информация, т.е. черпят се директно от първоизточника и са оригинални по естество. Изследователят трябва да отдели повече средства и самостоятелно да се потруди за събирането и преработката им, но резултатът винаги е по-точен и актуален от данните, събрани от вторичен източник. Някои възможни методи за получаване на първични данни са чрез:

- **Анкета**. Състои се от предварително подбрани въпроси, които биват зададени на определена целева група. Този метод се прилага от изследователи, частни лица, частни и обществени организации и от правителства (Mazhar et al., 2021). Пример за такова проучване е преброяването на населението, което в България се провежда онлайн чрез самостоятелен отговор на въпросите от страна на респондента, или от интервюиращ.
- **Наблюдение**. Това е най-често използваната практика, особено в проучвания, свързани с поведенческите науки. Този метод е особено подходящ при проучвания върху субекти, които по някаква причина не са в състояние да представят устно своите чувства. Пример за проучване чрез наблюдение е **преброяването** на броя автомобили, преминаващи през дадена улица за определен период от време (Mazhar et al., 2021).
- **Експеримент**. Изследователят задава предварително параметрите на средата и наблюдава ефектите ѝ върху даден обект. За да няма съмнение във верността на **информацията**, проучването трябва по възможност да се повтори, тестовата група да бъде представителна и резултатите да бъдат сравнени с контролна група, която не е била подлагана на експеримента.

Вторичните данни са информация, получена от предварително изготвени източници. За разлика от първичните данни, тук не участваме в процеса на проучване, а получаваме резултатите и изводите наготово. Така се пестят време и средства, но се рискува използването на неточна или остаряла информация. Те могат да са **вътрешни** или **външни,** в зависимост от това дали използваме информация от същата организация, за която правим проучването.

Вътрешни източници на данни са информационните системи, използвани от организацията за ежедневни операции, например CRM, ERP, счетоводни, финансови системи, системи за

контрол и мониторинг и др. Вътрешните източници не са публично достъпни. Организациите могат да ги използват за собствени анализи или да предоставят някои от тях за обработки.

Външни източници на данни са всички останали, които не са изготвени пряко за нашата организация. Тук влизат всички проучвания, статистики, публикации, статии и доклади на частни и юридически лица, правителствени и неправителствени организации и др. Често информацията, изготвена от правителствени организации, е безплатна и лесно достъпна за всички, но това не винаги важи за останалите изброени страни. Друго, което трябва да се вземе предвид, е областта на нашето проучване при избиране на подходящата платформа. Ако например изготвяме научен доклад, един полезен вторичен външен източник би бил **GOOGLE SCHOLAR**. Медицински лица използват също и **PUBMED**. Подобна платформа за икономисти е **IDEAS**, въпреки че за правилната икономическа оценка е нужен прегледът и на много други статистики и индекси.

Интернет източниците на данни са все по-предпочитан външен източник на данни поради предлагания богат избор от информация и удобния начин за достъп. Големият набор от непрекъснато увеличаващи се и усъвършенстващи се интернет услуги води до увеличаване на глобалния обем от данни в мрежата. Само за 2020-та година в интернет се генерира информация в размер на 64.2 ZB, като прогнозите за 2024-та година са тази стойност да стигне 147 ZB и да нараства с приблизително ~23% годишно (Taylor, 2023). Според Bankov (2017), интересът в областта на обработката и прилагането на информация, генерирана онлайн, се покачва значително поради нарастващия обем публикации в социалните мрежи.

Все повече фирми използват интернет канали за дигитализация на комуникацията и съхраняване на документи с цел повишаване на производителността и конкурентоспособността си на глобалния пазар (Sulova, 2021). Потенциалните предимства са значителни – чрез дигитализация на процеси, работещи с големи обеми данни, може значително да се съкрати времето за събиране, обработка и вземане на решения, базирани на тази информация. Освен това, преминаването от хартиен към софтуерен подход позволява автоматизиране на събирането на данни и подобряване на тяхното качество. Това от своя страна създава възможности за анализи в реално време, намаляване на рисковете, увеличаване на приходите, оптимизиране на разходите и други. (Kuyumdzhiev & Andreeva, 2023).

Приложението на иновативни технологии, каквито са Интернет на нещата (IoT), изкуствения интелект и анализи на Big Data в различни индустрии се изследва от много автори. Това се отнася и за големите езикови модели (Armyanova, 2022; Petrov et al., 2021; Parusheva & Aleksandrova, 2021). Въпреки че някои модели на изкуствен интелект не разкриват откъде са получили данните си за обучение, често е налична информация за това кои набори от данни или уеб обхождащи програми са били използвани (David, 2024). Доколкото е публично известно, те са били тренирани основно с информация от интернет. Данните в Интернет са специфични, динамични и постоянно променящи, което налага непрекъснатата нужда на големите езикови модели от актуална текстова информация, за да могат да изпълнят предназначението си като „…софтуерни (или хардуерни) системи, създадени от хората, които, при зададена комплексна цел, извършват действие във физическите или дигиталните измерения като възприемат околната си среда чрез събиране на данни, интерпретиране на събраната структурирана или неструктурирана информация, разсъждения върху знанията, или обработка на информацията, която е произлязла от тези данни и решава най-доброто действие/действия за да постигне тази цел“ (Angelova, Nisheva-Pavlova, Eskenazi and Ivanova, 2021).

Според Сълова (2021), основните източници на данни от интеренет ресурси са:

• Данните от уеб базирани и мобилни приложения, например тези за дигитална търговия и

резервации. Тези данни са структурирани и се съхраняват в бази от данни.

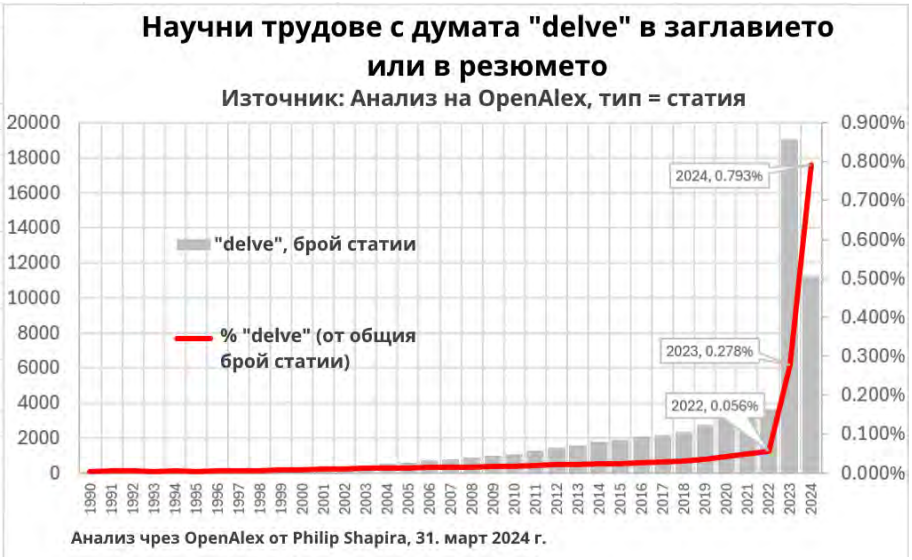
- Данни от приложения, работещи с технологията IoT. Тези данни се генерират в реално време на потоци. Могат да бъдат структурирани (числови стойности), или неструктурирани (аудио и видео записи, изображения и друг снимков материал).
- Данни от форуми, блогове и други уебсайтове, съдържащи потребителски коментари и ревюта.
- Данни от социални медии.
- Данни от имейл съобщения.
- Данни, генерирани при използване на уеб приложенията и сайтовете.

2. СЪВРЕМЕННИ ПРОБЛЕМИ, СВЪРЗАНИ СЪС СЪБИРАНЕТО НА ДАННИ

Въпреки изобилието от данни в днешно време и тяхната привидна достъпност, парадоксално се получава така, че работата на специалистите, занимаващи се с извличане, обработка и анализ на данни, се усложнява поради монополизирането на областта от технологичните гиганти и разпространението на генеративния изкуствен интелект. Това установява Robyn Speer, създателят на Python библиотеката *wordfreq*, която от 2016-та година предоставя достъп до оценки на честотата на използване на дадена дума на над 40 езика.

През септември 2024-та година заявява, че работата в областта на обработката на естествения език се превръща в безполезна и закрива дейностите си по проекта. Отправя критики към публичните платформи Reddit и Twitter поради това, че прекратяват достъпа до безплатните си публични API-та, с което възпират независимите изследвания върху неформалния начин на изказ онлайн и така фаворитизират OpenAI и Google. Упреква ги и в стигматизирането на събирането на текстова информация, защото използват чужди текстове за користни цели, най-вече за създаването на своите частни големи езикови модели.

Друго препятствие, което изтъква като съществено – модерните източници на лингвистични данни са „замърсени“ от генеративен изкуствен интелект и никой няма представа какъв е истинският човешки изказ след 2021-ва година. Това твърдение се потвърждава и от Philip Shapira, който установява, че ChatGPT неестествено завишава използването на някои думи в научните трудове, каквато е и думата “delve”.



Фигура 2. Използване на ключовата дума “delve” в заглавията и резюметата на научни публикации през годините
Източник: Shapira, 2024

Използването на тази дума се завишава през 2022-ра година, но значително нараства след 2023-та година, когато използването на ChatGPT става масово достъпно. Този феномен представлява значителна пречка в изследването на текстови източници онлайн, тъй като "delve" е само една от десет установени подобни думи, върху които ChatGPT има фиксация. Въпреки това, използването на определени думи не може да бъде точен критерий за намесата на изкуствен интелект в даден текст, тъй като е напълно възможно думата да е част от екзекутивния речник на автора. Robyn Speer твърди, че винаги е имало спам на думи в източниците на данни, използвани от *wordfreq*, но изкривяване на честотата на използване от такъв мащаб е невъзможно да бъде отсято, имайки се предвид и че генерираният текст е привидно написан от човешка ръка.

Подобно „замърсяване" на данните винаги трябва да бъде предварително изчистено, преди да се пристъпи към изваждането на заключения. За да се осигури качеството на данните, задължително трябва да се оцени тяхната актуалност, консистентност и обозримост на поставения проблем. Според Petrov, et al. (2020), новите изисквания за работа в дигитална среда изискват спазване на следните основни принципи при избора и съхранението на данни:

1. Съхраняване само на данни, които са от съществено значение за бизнеса и отговарят на променящите се изисквания на дигиталната среда.
2. Данните трябва да бъдат с високо качество, извлечени от надеждни източници, с гарантирана точност и прецизност.
3. Необходимо е запазването както на актуални, така и на исторически данни, тъй като и двата типа играят различни, но допълващи се роли в бизнес процесите.
4. Прилагане на гъвкави методи за интегриране на системи и внедряване на иновативни подходи, които осигуряват ефективно събиране и обработка на нови данни.
5. Изграждане на цялостна структура от метаданни, която улеснява управлението на данните и осигурява бърз и лесен достъп до необходимите извадки.

Сериозно предизвикателство поставя и използването на лични данни на физически лица като източник. В това число е всяка чувствителна информация, злоупотребата с която би нанесла вреда за дадено лице, каквито са медицинските досиета, информацията по банкови сметки и документите. Подобен риск представлява и всяко друго мероприятие, навлизащо в личното пространство на индивида, каквито са различните социологически проучвания и ползването на социални медии. Позволяването на достъп до информация на лица, които нямат основание да я използват, може да доведе до сериозни щети. Затова е добра практика да се следва подходът на минимизиране на данни. Например, една система, проследяваща броя души на една локация, не се нуждае от самоличността на даден потребител, а само да отчете неговото присъствие. Съществува и вариант, в който се предоставя допълнителна информация, с цел търговците да изготвят персонализирани реклами според демографски признаци и предишна история в дигиталното пространство, но преди това трябва да се поиска позволението на лицето (Stoyanova, Vasilev, Cristescu, 2021). Минимизирането на данни се прилага и с помощта на анонимизиране или псевдономизиране – два термина, използвани от европейското законодателство при създаването на Общия регламент относно защитата на данните (GDPR).

Псевдонимизацията, според определението на Европейския парламент, представлява обработка на лични данни по начин, който предотвратява свързването им с конкретен субект на данни без използването на допълнителна информация. Тази допълнителна информация трябва да се съхранява отделно и да бъде защитена чрез технически и организационни мерки, за да се гарантира, че данните не могат да бъдат свързани с идентифицирано или идентифицируемо физическо лице.

Тъй като псевдонимизираните данни могат да бъдат проследени обратно до индивида чрез допълнителната информация, те все още се считат за лични данни и попадат под регулациите на GDPR.

Анонимизацията е описана като процес, който, за разлика от псевдонимизацията, напълно заличава възможността да се проследи дадена информация обратно до източника ѝ. Така тези данни вече не попадат в обзора на GDPR и могат да бъдат използвани свободно за проучвания. Пример за анонимизация е групирането на годините на раждане на респондентите по периоди от анкета с цел да се предотврати изпъкването на определено лице и да бъде проследено. Подобно замаскиране на детайлите може да попречи на целите на изследването, но алтернативата е или да се приложи псевдонимизиране заедно с изискванията на GDPR, или данните въобще да не се използват.

Въпреки регулациите за неприкосновеност на личните данни, злоупотребите често влизат във полезрението на обществото поради нарушения. Подобен случай е скандалът със събирането на данни на потребители във Facebook и последвалото им използване от фирмата Cambridge Analytica с цел социален инжинеринг. Компанията използва Facebook като платформа за своето приложение "This Is Your Digital Life", което събира информация на потребител и неговите приятели. Така впоследствие става ясно, че от едва 270 000 души, използвали приложението, засегнатите са общо 87 милиона, въпреки че никога не са давали съгласието си на приятелите си или на Cambridge Analytica да използват и разпространяват личните им данни. Освен че информацията е събрана чрез манипулативни методи, тя по-късно се използва за профилиране и таргетиране на жертвите с цел по-ефективното прокарване на конкретни политически идеи. Последва съдебно производство за нелоялни практики срещу Meta, родителската компания на Facebook, при което заплаща $725 млн. долара като споразумение за уреждане на делото. Опасения за подобни злоупотреби с лични данни има и за платформата TikTok и практиките на родителската ѝ компания ByteDance. Остават неясни правилата за съгласие, както и начините на съхранение на потребителски данни от фирмата и за какви цели ги използва. Поради тази причина, някои държави забраняват използването на TikTok на правителствени устройства, а други забраняват приложението изцяло, включително и за обикновени граждани. Прави впечатление, че въпреки централата на ByteDance да е ситуиарана в Пекин, Китайската народна република е сред държавите, забранили TikTok. Вместо това, позволява негова стриктно контролирана версия, наричана Douyin, като по този начин единствено засилва съмненията към автентичността на компанията и нейните намерения.

ЗАКЛЮЧЕНИЕ

Основата на всяко едно добро проучване е в неговата точност и анализите, които се базират на данни. Правилното комбиниране на първични и вторични източници е ключово за изграждането на цялостна представа върху тематиката на изследвания въпрос. Развитието на технологиите безсъмнено налага един нов стандарт, насочен към дигитализация на процесите, без значение от сферата на дейност. Модерните изследователи и специалисти по обработка на данни трябва да се съобразяват с фактори, които не са имали тези размери само допреди няколко години, или въобще не са съществували. Огромните данни, които се генерират всеки ден, представляват предизвикателство за тяхното събиране, съхранение и обработка. Анализът на все по-големите масиви от данни изисква определени познания от страна на изследователите, както и специализирани технологии и ресурси. Единствено с комплексни решения и колаборация между технологичните компании, правителствата и потребителите могат да се спазят всички съвременни изисквания за качество на данните и етичност при събирането им.

REFERENCES / ИЗПОЛЗВАНА ЛИТЕРАТУРА

1. Сълова, С. (2021) Извличане на знания и анализ на данни от интернет източници. Варна: Наука и икономика.
2. Angelova, G., Nisheva-Pavlova, M., Eskenazi, A. and Ivanova, K. (2021) Role of education and research for artificial intelligence development in Bulgaria until 2030. *Mathematics and Education in Mathematics*. 50, (Apr. 2021), 71–82.
3. Armyanova, M. (2022) Artificial Intelligence and its Role in Modern Business. In Scientific Conference of the Department of General Economic Theory (pp. 184-191). University of Economics-Varna.
4. Bankov, B. (2017) Extracting Top Trends from Twitter Discussions in Bulgarian. Izvestia Journal of the Union of Scientists – Varna. Economic sciences series, 2, pp. 254-259.
5. David, E. (2024) *OpenAI's news publisher deals reportedly top out at $5 million a year, The Verge*. Available at: https://www.theverge.com/2024/1/4/24025409/openai-training-data-lowball-nyt-ai-copyright (Accessed: 09 November 2024).
6. geeksforgeeks.org (2024) *Sources of data collection: Primary and secondary sources* (2024) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/sources-of-data-collection-primary-and-secondary-sources/ (Accessed: 21 September 2024).
7. geeksforgeeks.org, (2024) *Different sources of data for data analysis* (2024) GeeksforGeeks. [Online] Available at: https://www.geeksforgeeks.org/different-sources-of-data-for-data-analysis/ (Accessed: 21 September 2024).
8. geeksforgeeks.org, (2024) *Methods of data collection: Statistics* (2024) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/methods-of-data-collection/ (Accessed: 21 September 2024).
9. Kuyumdzhiev, I., & Andreeva, A. (2023) Digital transformation of the general ledger process in universities: problems and possible solutions. Munich Personal RePEc Archive. *Uni-muenchen.de*. [online] doi:https://mpra.ub.uni-muenchen.de/116591/1/MPRA_paper_116591.pdf.
10. Mazhar SA, Anjum R, Anwar AI, Khan AA. (2021) Methods of data collection: A fundamental tool of research, *Journal of Integrated Community Health*. Vol. 10, No 21, pp. 6-10, DOI: https://doi.org/10.24321/2319.9113.202101.
11. Mesly, O. (2015) *Creating models in psychological research*. Etats-Unis: Springer press.
12. Parusheva, S., & Aleksandrova, Y. (2021) Digital technologies and tools-drivers of digitalization in construction. Izvestia Journal of the Union of Scientists-Varna. Economic Sciences Series, 10(1), 63- 71.
13. Petrov, P., Sulova, S., Radev, M., Aleksandrova, Y., Stoyanova, M., Mileva, L. and Yankov, P., (2020), Digitalization of business processes in construction and logistics, Publishing house "Knowledge and business" Varna, https://EconPapers.repec.org/RePEc:kab:monogr:8.
14. Petrov, P., Vasilev, J., Petrova, S., & Mileva, L. (2021) Technological organization in big data analysis with Apache Kudu analytical tool. Izvestia Journal of the Union of Scientists-Varna. Economic Sciences Series, 10(2), 31-42.
15. Shapira, P. (2024) *Delving into 'delve'*, Benedictine University Library. Available at: https://pshapira.net/2024/03/31/delving-into-delve/ (Accessed: 29 September 2024).
16. Speer, R. (2024) *Why wordfreq will not be updated*, GitHub. Available at: https://github.com/rspeer/wordfreq/blob/master/SUNSET.md (Accessed: 29 September 2024).
17. Stoyanova, M., Vasilev, J., Cristescu, M. (2021) Big Data in Property Management. Applications of Mathematics in Engineering and Economics: Proceedings of the 46th Conference on Applications of Mathematics in Engineering and Economics (AMEE '20), 7 – 13 June 2020, Sofia, Bulgaria, Melville,

NY: AIP Publ., Vol. 2333, № 1, 070001-1 - 070001-7.

18. Sulova, S. (2021) A conceptual model for the organization and storage of metadata for data from internet sources. [online] (2), pp.3–14. Available at: https://www.researchgate.net/publication/363309614_A_conceptual_model_for_the_organization_and_storage_of_metadata_for_data_from_internet_sources.
19. Taylor, P. (2023) *Data Growth Worldwide 2010-2025*, Statista. Available at: https://www.statista.com/statistics/871513/worldwide-data-created/ (Accessed: 09 November 2024).
20. Wagh, S. (2024) Research guides: Public Health Research Guide: Primary & Secondary Data Definitions, Primary & Secondary Data Definitions - Public Health Research Guide - Research Guides at Benedictine University Library. Available at: https://researchguides.ben.edu/c.php?g=282050&p=4036581 (Accessed: 28 September 2024).