OTWARTE DANE A SZTUCZNA INTELIGENCJA

RAPORT Z BADAŃ WSTĘPNYCH

DOMINIKA KACZOROWSKA-SPYCHALSKA

SEBASTIAN KONDRACKI

Raport przygotowany przez

Grupę Roboczą ds. Sztucznej Inteligencji Podgrupa ds. badań, innowacyjności i wdrożeń

we współpracy z

Centrum Inteligentnych Technologii Wydziału Zarządzania UŁ,

Deviniti

GRAI
GRUPA ROBOCZA
DS. SZTUCZNEJ INTELIGENCJI

CENTRUM INTELIGENTNYCH TECHNOLOGII
Wydział Zarządzania
Uniwersytet Łódzki

DEVINITI

POGLĄDY WYRAŻONE W TYM DOKUMENCIE SĄ ODPOWIEDZIALNOŚCIĄ AUTORÓW I NIE MUSZĄ KONIECZNIE ODZWIERCIEDLAĆ STANOWISKA POLSKIEGO RZĄDU

.

Warszawa-Łódź, 2022

Spis treści

AUTORZY **3**

ZAPROSZENIE EKSPERCI **5**

WSTĘP **8**

OTWARTE DANE – UJĘCIE DEFINICYJNE **11**

PODSTAWY PRAWNE **14**

O RAPORCIE **16**

KONTEKST BADAWCZY 16

METODYKA BADAŃ 20

CHARAKTERYSTYKA ZBIOROWOŚCI 21

OTWARTE DANE A SZTUCZNA INTELIGENCJA **23**

PORTAL OTWARTE DANE – DANE.GOV.PL 23

WYBRANE PRZYKŁADOWE ZAGRANICZNE PORTALE DOSTARCZAJĄCE OTWARTYCH DANYCH 27

OTWARTE DANE A AI W OPINII PRZEDSIĘBIORSTW **32**

POTRZEBY ORGANIZACJI W OBSZARZE OTWARTYCH DANYCH 32

GOTOWOŚĆ UDOSTĘPNIANIA DANYCH PRZEZ PRZEDSIĘBIORCÓW 33

BARIERY I OGRANICZENIA SPOWALNIAJĄCE DYNAMIKĘ ROZWOJU OTWARTYCH DANYCH W POLSCE, W TYM W KONTEKŚCIE AI 36

WNIOSKI I REKOMENDACJE **39**

BIBLIOGRAFIA **40**

Autorzy

dr Dominika Kaczorowska-Spychalska

Dyrektor Centrum Inteligentnych Technologii na Wydziale Zarządzania Uniwersytetu Łódzkiego. Pełni funkcję Lidera podgrupy Badań, Innowacyjności i Wdrożeń Grupy Roboczej ds. Sztucznej Inteligencji przy Kancelarii Prezesa Rady Ministrów. Zajmuje się w szczególności sztuczną inteligencją (AI) i Internetem Rzeczy (IoT) oraz ich implikacją w biznesie. Pasjonuje ją wpływ tych technologii na zachowania człowieka (Homo Cyber versus Homo Roboticus) oraz zakres i charakter interakcji między nim a technologią, w tym problematyka digital ethics. Jest autorką (w części współautorką) licznych publikacji wydanych zarówno w wydawnictwach polskich, w tym czasopismach branżowych, jak i zagranicznych. Była w grupie specjalistów zaangażowanych w prace nad przygotowaniem „Założeń do strategii AI w Polsce", a w 2020 roku znalazła się wśród laureatek rankingu Perspektyw Top 15 Women in 5G. Członek Rady Programowej Centrum Etyki Technologii i Fundacji Forum Inteligentnego Rozwoju. W 2022 roku została laureatką konkursu Centrum Inteligentnego Rozwoju – Naukowiec Przyszłości, w kategorii *Badania przyszłości*.

Sebastian Kondracki

Szef innowacji w Deviniti, jednej z najdynamiczniej rozwijających się firm tworzących oprogramowanie w Polsce. Członek podgrupy Badań, Innowacyjności i Wdrożeń Grupy Roboczej ds. Sztucznej Inteligencji przy Kancelarii Prezesa Rady Ministrów. Ekspert ds. transformacji cyfrowej zwłaszcza w obszarze jej głównych akceleratorów, jak: procesy innowacyjne i szeroko rozumiana sztuczna inteligencja. Autor książki „Python i AI dla e-commerce", a także wykładowca na studiach podyplomowych w Akademii Leona Koźmińskiego i Wyższej Szkole Bankowej we Wrocławiu. Aktywny badacz wpływu "otwartości" (open data, open science, open source) i ich wpływu na rozwój innowacyjności i kreatywności w społeczeństwie i biznesie. Współtwórca projektu SpeakLeash/ˈspix.lɛʂ/a.k.a.Spichlerz - zestawu narzędzi do budowy pierwszego w Polsce tak obszernego (>1 TB) zróżnicowanego zbioru danych wysokiej jakości wraz z bibliotekami umożliwiającymi łatwy dostęp i pełną otwartą replikację danych lub ich analizę. Specjalizuje się w AI: NLP (Natural Language Processing), NLG (Natural Language Generation), Generative AI, Explainable Artificial Intelligence (XAI) w biznesie, a w szczególności w e-commerce. Ekspert w zakresie testowania i weryfikacji pomysłów biznesowych, lean startup, zarządzania projektami eksploracyjnymi. Autor licznych publikacji branżowych, w tym dla magazynu komputerowego CHIP, CHIP Special i Computer World.

dr Maciej Adamiak

Jest doktorem nauk ścisłych i przyrodniczych w dyscyplinie nauki o Ziemi i środowisku. W pracy naukowej oraz zawodowo zajmuje się zastosowaniem uczenia maszynowego w teledetekcji, fotointerpretacji oraz analizie przestrzennej. Prezes zarządu oraz CSO w spółce ReasonField Lab

dr inż. Krzysztof Rychlicki-Kicior

Prezes zarządu i Chief Scientific Officer w firmie tworzącej oprogramowanie – Makimo sp. z o.o., a także adiunkt i dziekan Wydziału Technologii Informatycznych na Akademii Ekonomiczno – Humanistycznej w Warszawie. Autor książek informatycznych, wykładowca, popularyzator sztucznej inteligencji. W pracy naukowej zajmuje się analizą obrazów medycznych z wykorzystaniem metod głębokiego nauczania, a także cyfrowym przetwarzaniem sygnałów w kontekście nagrań muzycznych.

Piotr Zawadzki

Radca prawny i rzecznik patentowy w kancelarii Bird&Bird. W pracy zawodowej wspiera klientów w sprawach z zakresu prawa patentowego, prawa autorskiego, regulacji e-commerce, prawa reklamy, ochrony tajemnic przedsiębiorstwa, a także w prawie danych osobowych oraz regulacjach dotyczących danych nieosobowych. Jest także autorem lub współautorem prawie 40 publikacji dotyczących opisanych powyżej tematów, członkiem organizacji branżowych zajmujących się własnością intelektualną oraz wykładowcą akademickim.

Szanowni Państwo,

Przekazujemy w Wasze ręce raport będący zestawieniem wyników wstępnych badań nad wykorzystaniem danych w dalszym rozwoju AI w polskich przedsiębiorstw, które koncentrują się na opracowywaniu i wytwarzaniu nowych technologii lub usług dla swoich klientów. Sztuczna inteligencja stała się dla nich nieodzownym narzędziem pozwalającym kreować nowe modele biznesowe, innowacyjne produkty i usługi oraz walczyć o swoją pozycję konkurencyjną. Czego jednak oczekują od dostępnych na rynku ogromnych zbiorów danych? Czy potrafią z nich umiejętnie korzystać? Czy są skłonne do współdzielenia się własnymi danymi? Czy staną się w tym zakresie katalizatorem zmian dla swoich klientów? Pytań jest wiele, a postępująca transformacja cyfrowa wymaga od nich coraz większej dojrzałości w tym zakresie.

Niniejsza publikacja jest efektem prac w ramach Podgrupy ds. Badań, Innowacyjności i Wdrożeń Grupy Roboczej ds. Sztucznej Inteligencji. Prezentowane w raporcie zagadnienia wpisują się w trwającą obecnie debatę nad otwartymi danymi i ich rolą w dalszym rozwoju AI, a także szerokim spektrum wynikających z tego wyzwań i konsekwencji, zarówno o charakterze biznesowym, społecznym, jak i naukowym. Opracowanie stanowi próbę wstępnej identyfikacji czynników determinujących istotę i znaczenie otwartych danych w kontekście sztucznej inteligencji. Wyniki badań zostały wzbogacone komentarzami ekspertów reprezentujących interdyscyplinarne obszary rynku, co pozwala spojrzeć na omawiane zagadnienia w sposób holistyczny. Raport nakreśla także szerokie spektrum możliwości i ram dalszych badań oraz wynikających z nich rekomendacji.

Mamy nadzieję, że materiał, jaki przekazujemy w Państwa ręce stanie się źródłem kolejnych inspiracji związanych zarówno z problematyką sztucznej inteligencji, jak i otwartych danych, intensyfikując kolejne badania, syntezy i prace podejmowane w tym obszarze.

Z wyrazami szacunku,

Autorzy

Wstęp

Zaawansowane technologie cyfrowe, w tym zwłaszcza sztuczna inteligencja (*Artificial Intelligence - AI*), coraz bardziej stają się integralną częścią naszego życia, zmieniając sposób w jaki podmioty rynkowe podejmują decyzje i wchodzą w interakcję ze swoimi zewnętrznymi interesariuszami (np. pracownikami, klientami, społeczeństwem)[1]. Przyciągają uwagę i rozbudzają wyobraźnię, stając się nośnikiem fundamentalnych zmian. Pozwalają na kreację nowych modeli biznesowych, wdrażanie nowych narzędzi i strategii konkurowania, optymalizację dotychczasowych procesów czy kształtowanie się nowych potrzeb, oczekiwań, zachowań i postaw, łącząc nas w sieć cyfrowych powiązań, interakcji i zależności. Generowany w wyniku tego strumień danych pozwala na odkrycie szeregu wzorców i zależności Wszystko dzieje się wokół nas i jednocześnie poza nami, stanowiąc nowe cyfrowe otoczenie[2]. Dzięki temu poszczególne gospodarki, przedsiębiorstwa i społeczeństwo mogą ewoluować od analogowej przeszłości, poprzez cyfrową teraźniejszość, ku wyzwaniom autonomicznej przyszłości[3].

U podstaw tego procesu niewątpliwie leżą dane, postrzegane obecnie jako jedna z najważniejszych kategorii zasobów, którymi dysponujemy. Przeszliśmy bowiem od rewolucji informacyjnej, w której pojawiły się dane, do rewolucji danych pochodzących z wielu źródeł, obiektywnych i dostępnych w czasie rzeczywistym[4], czego efektem jest przepływ w każdej sekundzie ich ogromnych ilości pomiędzy niemal nieograniczoną liczbą nadawców i odbiorców. Świat nigdy nie zbierał bowiem i nie przechowywał tak dużych zbiorów danych jak dzisiaj. Co więcej, ich zróżnicowanie, ilość i tempo generowania kolejnych rosną w niewyobrażalny tempie[5] we wszystkich sferach naszej aktywności. Wielkości analogowe, jak chociażby: parametry fizyczne, chemiczne, obrazy, parametry procesów środowiskowych i przemysłowych, dotyczące zdrowia człowieka, stanu infrastruktury przedsiębiorstwa, bezpieczeństwa są digitalizowane, strukturyzowane, katalogowane i gromadzone[6]. W efekcie postępująca danetyzacja pozwala na przekładanie wszelkich elementów rzeczywistości na dane, już nie tylko na potrzeby konkretnego, zaplanowanego procesu czy projektu, ale ze względu na możliwość ich wykorzystania

[1] A.Kaplan, M.Haenlein. (2019), *Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence*, „Business Horizons", vol. 62, no. 1, s. 15-25
[2] D. Kaczorowska-Spychalska, Ł.Sułkowski (2021), *Determinants of the adoption of AI wearables - practical implications for marketing*, „Human Technology", vol. 17, no 3, s. 294-320.
[3] B.Gregor, D. Kaczorowska-Spychalska (2020), *Technologie cyfrowe w biznesie. Przedsiębiorstwa 4.0 a sztuczna inteligencja*, PWN, Warszawa, s. 9.
[4] A.Poniewierski (2020), *SPEED. Bez granic w cyfrowym świecie*, www.speednolimits.com, Warszawa s. 54.
[5] A.Castrounis, (2019), *AI for people and business,* O'Reilly Media, Inc, USA, s.89.
[6] R.Romaniuk (2020), *Systemy informatyczne jako fundament przedsiębiorstwa 4.0* [w:] B.Gregor, D. Kaczorowska-Spychalska, *Technologie cyfrowe w biznesie. Przedsiębiorstwa 4.0 a sztuczna inteligencja*, PWN, Warszawa, s.19.

w przyszłości w sposób, o którym na etapie ich gromadzenia nawet nie jesteśmy w stanie pomyśleć[7]. Nie wystarczy bowiem jedynie je posiadać, trzeba je rozumieć i umieć wykorzystać w praktyce. To właśnie dzięki temu, ale i poprzez to, jesteśmy w stanie budować swoją przewagę. Bowiem najbardziej perspektywiczną strategią dla nowoczesnych gospodarek staje się zarządzanie danymi jako wspólnym zasobem, o charakterystyce infrastruktury (*środek dla wielu aktorów i do wielu celów*)[8].

Zgodnie z raportem *Open Data Maturity 2021*, Polska znalazła się na czwartym miejscu wśród państw zaliczanych do kategorii: *Trendsetters*[9]. To grupa krajów, które realizują obecnie bardzo zaawansowaną politykę otwartych danych na wszystkich szczeblach swojej administracji. Udostępniają one szerokie spektrum danych o wysokim poziomie jakości, dostosowanych do potrzeb użytkowników, inicjują i wspierają przy tym powstawanie różnych ekosystemów otwartych danych, charakteryzujących się wysokim poziomem interakcji i ponownego ich wykorzystania. Proces ten podlega pomiarowi z zastosowaniem metodologii oceny wpływu w różnych dziedzinach[10]. Wśród kluczowych zaleceń dla dalszej polityki w zakresie otwartych danych, dla tej grupy państw, znalazły się między innymi kwestie związane z podnoszeniem ilości i jakości udostępnianych danych (*np. ich walidacja, konwersja danych na różne alternatywne formaty, lepsze zrozumienie potrzeb odbiorców danych, umieszczanie linków do źródeł danych udostępnianych w czasie rzeczywistym, umożliwienie użytkownikom opcji komentowania i oceny zestawów danych, z możliwością ich umieszczenia w algorytmach wyszukiwania*) i ich prezentacji, pogłębiania współpracy w różnymi grupami interesariuszy (*np. zespoły otwartych danych, uniwersytety, instytucje badawcze*) w celu opracowania ram i wskaźników pomiaru i oceny wpływu czy konsolidacja i wsparcie ekosystemów otwartych danych oraz optymalizacja strategii w zakresie otwartych danych.

W świecie, w którym trwa intensywna i nieustająca debata nad możliwościami udostępniania danych, z których obecnie wciąż wiele jest marnowanych, a sztuczna inteligencja staje się integralną częścią

---

[7] Ł. Iwasiński (2016), Społeczne zagrożenia danetyzacji rzeczywistości, [w:] M. Przastek-Samokowa, B. Sosińska-Kalata, Z. Wiorogórska (red.), *Nauka o informacji w okresie zmian. Informatologia i humanistyka cyfrowa*, Wydawnictwo SBP, Warszawa, s. 135-146.

[8] B. Wawrzyniak, M. Musidłowska, J. J. Zygmuntowski (2022), *Uwolnić potencjał danych. Zarządzanie danymi jako zasobem współdzielonym*, KPRM, Warszawa, s. 11.

[9] *Raport dokonuje oceny dojrzałości cyfrowej poszczególnych krajów biorąc pod uwagę cztery kryteria: realizowaną politykę w zakresie otwartych danych, ocenę funkcji jakie posiadają krajowe portale otwartych danych, w tym tych umożliwiających użytkownikom dostęp do danych i wsparcie zachodzących tam interakcji, wpływ w oparciu o działania mające na celu monitorowanie i pomiar ponownego wykorzystania otwartych danych, a także jakość (meta)danych*. D. van Hesteren, L. van Knippenberg (2021), *Open Data Maturity Report*, Luxembourg: Publications Office of the European Union, s. 3.

[10] D. van Hesteren, L. van Knippenberg (2021), *Open Data Maturity Report*, Luxembourg: Publications Office of the European Union, s. 94.

naszego życia, uzasadnione wydaje się założenie, że wszystkie podmioty rynkowe powinny zacząć coraz bardziej myśleć o sobie w kategorii kreatorów danych i ich analityków, niezależnie od ich głównego sektora operacyjnego. Głębokie osadzenie swojej działalności w perspektywie danych, które generujemy i zmiana dotychczasowego sposobu myślenia w tym zakresie, wydaje się coraz bardziej niezbędne do czerpania zysków przez ludzi i firmy[11], branże, sektory, grupy społeczne, poszczególne gospodarki i społeczeństwa. Całkowita wolność w zakresie dostępu do danych powinna jednak dotyczyć danych publicznych czy danych pochodzących ze środowiska naturalnego człowieka. Dane firm, szczególnie sektora MSP czy też dane dotyczące zdrowia obywateli, ze względu na swoją specyfikę, powinny być możliwe do wykorzystania w oparciu o wypracowane podejście danego sektora[12]. Tworząc bowiem technologie i je wykorzystując, możemy w istotnym stopniu zmieniać świat, aktywnie się do niego adaptując i/lub kreując go. Wraz z tą potęgą spada na nas jednak coraz większa odpowiedzialność[13].

[11] A. Castrounis, (2019), *AI for people and business,* O'Reilly Media, Inc., USA, s. 11.
[12] M. Borowik, L. Maśniak, R. Kroplewski, H. Romaniec (2017) *Przemysł + Gospodarka oparta o dane*, Ministerstwo Cyfryzacji, https://www.gov.pl/web/cyfryzacja/gospodarka-oparta-o-dane-przemysl. pdf [dostęp: 28.10.2022].
[13] E. Brynjolfsson, A. McAffe (2016), *The second machine age*, John Wiley & Sons, New York-London, s. 362.

Pomimo dużej popularności otwartych danych (m.in. proces ich rozwoju został uwzględniony w programach celowych dla rozwoju sztucznej inteligencji w Polsce[14]) istnieje pewna nieścisłość pomiędzy definicjami występującymi w literaturze zajmującej się tzw. instytucjami prawno-autorskimi (np. otwartymi licencjami), a definicją legalną zawartą w przepisach prawa[15]. Zanim jednak zajmiemy się tą nieścisłością, spójrzmy na najbardziej popularną definicję otwartych danych, zawartą na stronie WWW „Open Definition" (http://opendefinition.org/) pod opieką międzynarodowego stowarzyszenia Open Knowledge Foundation (OKF). Zatem ***otwarte dane (ang. open data, OD) według OKF to dane, które mogą być swobodnie wykorzystywane, modyfikowane i udostępniane przez każdego w dowolnym celu***[16]. Należy zauważyć, że OKF nie skupia się na źródle pochodzenia danych (rząd, ośrodki naukowe i akademickie czy biznes), a wyłącznie definiuję swobodę ich używanie.

Jednak na rozwój otwartych danych mają istotny wpływ trzy nurty:

- idea udostępniania bezpłatnych wyników badań naukowych,
- otwierania dostępu do danych przetwarzanych przez instytucje publiczne i rządowe,
- „zwrotu" danych samodzielnie generowanych przez użytkowników w aplikacjach internetowych.

Dlatego też możemy zdefiniować inne podtypy otwartych danych m. in.: **otwarte dane rządowe** (ang. *open government data, OGD*), **otwarte dane naukowe** (ang. *open scientific data*) czy też **otwarte dane badawcze** (ang. *open research data*). Należy nadmienić, że **otwarte dane rządowe** zaczynają się bardzo dynamicznie rozwijać, ponieważ doskonale wpisują się w potrzeby coraz bardziej cyfrowego społeczeństwa, w tym wsparcia publicznego nadzoru nad rządami, zmniejszenia korupcji, większej przejrzystości wydatków publicznych itp. Według OECD (Organizacja Współpracy Gospodarczej i Rozwoju) **otwarte dane rządowe** (ang. *open government data, OGD*) **są przejawem swoistej filozofii (i w coraz większym stopniu zbiorem zasad), która promuje przejrzystość, odpowiedzialność**

[14] *Polityka dla rozwoju sztucznej inteligencji w Polsce od roku 2020 (praca zbiorowa) (2020),* załącznik do uchwały nr 196 Rady Ministrów z dnia 28 grudnia 2020 r. (poz. 23), s. 3.
[15] *Ustawa z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego* (Dz. U. 2021 poz. 1641).
[16] http://opendefinition.org/ [dostęp: 11.08.2022].

i tworzenie wartości poprzez udostępnianie danych rządowych wszystkim zainteresowanym[17]. Rozwój OGD jest na tyle dynamiczny w porównaniu z innymi źródłami otwartych danych np. nauką, kulturą czy biznesem, że następuje swoiste „przejęcie" nazwy tj. traktowania otwartych danych jako wyłącznie otwartych danych pochodzących z organów administracji publicznej. Taka właśnie nieścisłość pojawiła się w definicji otwartych danych w *Ustawie z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego (Dz. U. 2021 poz. 1641)*. Zgodnie z nią **otwarte dane są definiowane jako „informacje sektora publicznego udostępniane lub przekazywane w postaci elektronicznej, bezwarunkowo lub z uwzględnieniem warunków, o których mowa w rozdziale 3 tej ustawy, kompletne, aktualne, w wersji źródłowej, w otwartym i niezastrzeżonym formacie przeznaczonym do odczytu maszynowego, które są przeznaczone do bezpłatnego ponownego wykorzystywania na tych samych zasadach dla każdego użytkownika, bez konieczności potwierdzania tożsamości przez użytkownika."**[18]

Należy dodać, że rozdział 3, wspomniany w definicji, określa dodatkowe warunki jak chociażby, kiedy udostępniane informacje mają charakter utworu, martyrologii lub zawierają godło, barwy i hymn Rzeczypospolitej Polskiej lub podmiot udostępniający wymaga dodatkowych warunków znanych ze standardowych otwartych licencji (np. obowiązek poinformowania o źródle, czasie wytworzenia i pozyskania informacji sektora publicznego)[19].

Nieścisłość opisana wyżej ma charakter wyłącznie definicyjny i nie ma wpływu na rozwój otwartych danych w Polsce. Przykładowo portal https://dane.gov.pl/pl funkcjonuje na podstawie:

- ustawy z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego (Dz. U. poz. 1641),
- rozporządzenia Rady Ministrów z dnia 12 marca 2014 r. w sprawie Centralnego Repozytorium Informacji Publicznej (Dz. U. z 2014 r., poz. 361, z późn. zm.),

a jego przebudowa jest dofinansowana z projektów: *"Otwarte Dane - dostęp, standard, edukacja"* oraz *„Otwarte dane plus"* POPC[20] publikuje zarówno dane pochodzące z administracji publicznej, jak i podmiotów prywatnych.

[17] https://www.oecd.org/gov/digital-government/open-government-data.htm [dostęp: 17.08.2022]
[18] Ustawie z dnia 11 sierpnia 2021 r. O otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego (Dz. U. 2021 poz. 1641).
[19] *Szersza analiza definicji legalnej otwartych danych zawarta jest w części: Podstawy prawne*.
[20] https://dane.gov.pl/pl/page/o-serwisie?lang=pl&rev=-1 [dostęp: 18.08.2022].

Zatem z racji semantycznej różnicy pomiędzy definicjami zawartymi w literaturze a definicją legalną zgodną z ustawą w dalszej części raportu będziemy używać terminu:

- **otwartych danych** dla danych, które mogą być swobodnie wykorzystywane, modyfikowane i udostępniane przez każdego w dowolnym celu[21], niezależnie od źródła pochodzenia,

- **otwartych danych (OGD)** lub **otwartych publicznych danych** tj. **danych zgodnych z definicją zawartą** w ustawie z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego (Dz. U. 2021 poz. 1641),

- **otwartych danych z biznesu** lub **otwartych danych (OBD)** dla danych, które mogą być swobodnie wykorzystywane, modyfikowane i udostępniane przez każdego w dowolnym celu[22] (ewentualnie rozszerzone o warunki zawarte w standardowych otwartych licencjach np. MIT[23]) jednak pochodzących od prywatnych podmiotów biznesowych.

[21] http://opendefinition.org/ [dostęp: 11.08.2022].
[22] Tamże.
[23] https://opensource.org/licenses/MIT [dostęp: 11.08.2022].

Podstawy prawne

Głównym aktem prawnym w Polsce definiującym otwarte dane i zasady, tryb udostępniania i przekazywania informacji sektora publicznego w celu ponownego ich wykorzystywania jest ***ustawa z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego (Dz. U. poz. 1641)***. Ustawa weszła w życie 8 marca 2022 roku, w celu dostosowania polskiego prawa do przepisów Unii Europejskiej wdrażając dyrektywę Parlamentu Europejskiego i Rady (UE) 2019/1024 z dnia 20 czerwca 2019 r. w sprawie otwartych danych i ponownego wykorzystywania informacji sektora publicznego (Dz. Urz. UE L 172 z 26.06.2019, str. 56)[24]. Dodatkowo, główny akt prawny związany z otwartymi danymi zmienia ustawy:

- ustawę z dnia 17 maja 1989 r. – Prawo geodezyjne i kartograficzne,
- ustawę z dnia 29 czerwca 1995 r. o statystyce publicznej,
- ustawę z dnia 21 listopada 1996 r. o muzeach,
- ustawę z dnia 20 czerwca 1997 r. – Prawo o ruchu drogowym,
- ustawę z dnia 20 sierpnia 1997 r. o Krajowym Rejestrze Sądowym,
- ustawę z dnia 6 września 2001 r. o dostępie do informacji publicznej,
- ustawę z dnia 17 lutego 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne,
- ustawę z dnia 8 września 2006 r. o Państwowym Ratownictwie Medycznym, ustawę z dnia 24 września 2010 r. o ewidencji ludności,
- ustawę z dnia 20 lipca 2017 r. – Prawo wodne,
- ustawę z dnia 24 listopada 2017 r. o imprezach turystycznych i powiązanych usługach turystycznych,
- ustawę z dnia 6 marca 2018 r. o Centralnej Ewidencji i Informacji o Działalności Gospodarczej i Punkcie Informacji dla Przedsiębiorcy,
- ustawę z dnia 6 marca 2018 r. o zasadach uczestnictwa przedsiębiorców zagranicznych i innych osób zagranicznych w obrocie gospodarczym na terytorium Rzeczypospolitej Polskiej oraz uchyla się ustawę z dnia 25 lutego 2016 r. o ponownym wykorzystywaniu informacji sektora publicznego.

[24] https://eur-lex.europa.eu/legal-content/PL/ALL/?uri=CELEX%3A32019L1024 [dostęp: 23.08.2022].

Ustawa nie tylko definiuje same otwarte dane, ale także pojęcia ściśle powiązane, jak chociażby: **dane badawcze**, **dane dynamiczne**, **dane o wysokiej wartości**, **dane prywatne**, **format otwarty** i **format przeznaczony do odczytu maszynowego**. Specyfikuje tzw. **podmioty zobowiązane** do udostępniania lub przekazywania informacji sektora publicznego w celu ponownego wykorzystywania. Instytucja podmiotów zobowiązanych nie tylko zawiera jednostki sektora finansów publicznych w rozumieniu przepisów ustawy z dnia 27 sierpnia 2009 r. o finansach publicznych (Dz. U. z 2021 r. poz. 305, 1236 i 1535), ale także podmioty, które nie są związane z sektorem publicznym np. działające w charakterze przewoźników lotniczych. Jednak wykluczając podmioty m.in.: jednostki publicznej radiofonii i telewizji w rozumieniu przepisów ustawy z dnia 29 grudnia 1992 r. o radiofonii i telewizji (Dz. U. z 2020 r. poz. 805) oraz Polskiej Agencji Prasowej S.A.

Dodatkowo ustawa definiuje szczegółowo zasady udostępniania i przekazywania informacji sektora publicznego w celu ponownego ich wykorzystywania oraz warunki ponownego wykorzystania. Zwłaszcza w przypadku informacji z sektora publicznego, mających cechy utworu lub przedmiotu praw pokrewnych w rozumieniu przepisów ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych lub stanowiących bazę danych w rozumieniu przepisów ustawy z dnia 27 lipca 2001 r. o ochronie baz danych lub objętych prawami do odmian roślin w rozumieniu przepisów ustawy z dnia 26 czerwca 2003 r. o ochronie prawnej odmian roślin.

Ustawa z 2021 r. stanowi kontynuację obowiązujących w Polsce już wcześniej regulacji w zakresie ponownego wykorzystywania informacji sektora publicznego („ISP"). Nowy akt (podobnie zresztą jak jego poprzednik) stanowi wdrożenie unijnych dyrektyw dotyczących omawianej tematyki (obecnie jest to dyrektywa (UE) 2019/1024). Wprowadza tym samym do polskiego systemu prawnego, zasady i standardy obowiązujące w Unii Europejskiej. Przepisy o ponownym wykorzystywaniu ISP, wraz z ustawą z 6 września 2001 r. o dostępie do informacji publicznej, stanowią trzon regulacji pozwalających na udostępnianie danych będących w posiadaniu organów publicznych (i innych podmiotów zobowiązanych). Do naczelnych zasad tego systemu należy, przede wszystkim, reguła jawności. Wszelkie odstępstwa od niej należy więc traktować jako wyjątki (a zatem intepretować w sposób zawężający).

Nie mniej ważny jest także prawny wymóg stworzenia powszechnie dostępnego systemu teleinformatycznego służącego do udostępniania ISP (zrealizowany przez powstanie portalu dane.gov.pl). Ustawa przewiduje jednak i inne rozwiązania organizacyjne oraz informatyczne (np. interfejsy programistyczne aplikacji – API), dzięki którym dostęp do danych ma być znacznie łatwiejszy (w porównaniu z „tradycyjnym" trybem wnioskowym). I to właśnie od zastosowania nowoczesnych rozwiązań pozwalających na szybki, bezwnioskowy, a – w razie konieczności – stały dostęp do zasobów, zależeć będzie dynamika rozwoju systemów AI korzystających z danych pozostających pod kontrolą sektora publicznego. Ustawa z 2021 r. daje bowiem narzędzia pozwalające na stworzenie systemu, który nie będzie generował faktycznych i organizacyjnych przeszkód dla korzystania z danych. Sektor publiczny powinien więc z niego efektywnie korzystać.

Piotr Zawadzki
Radca prawny i rzecznik patentowy w kancelarii Bird&Bird

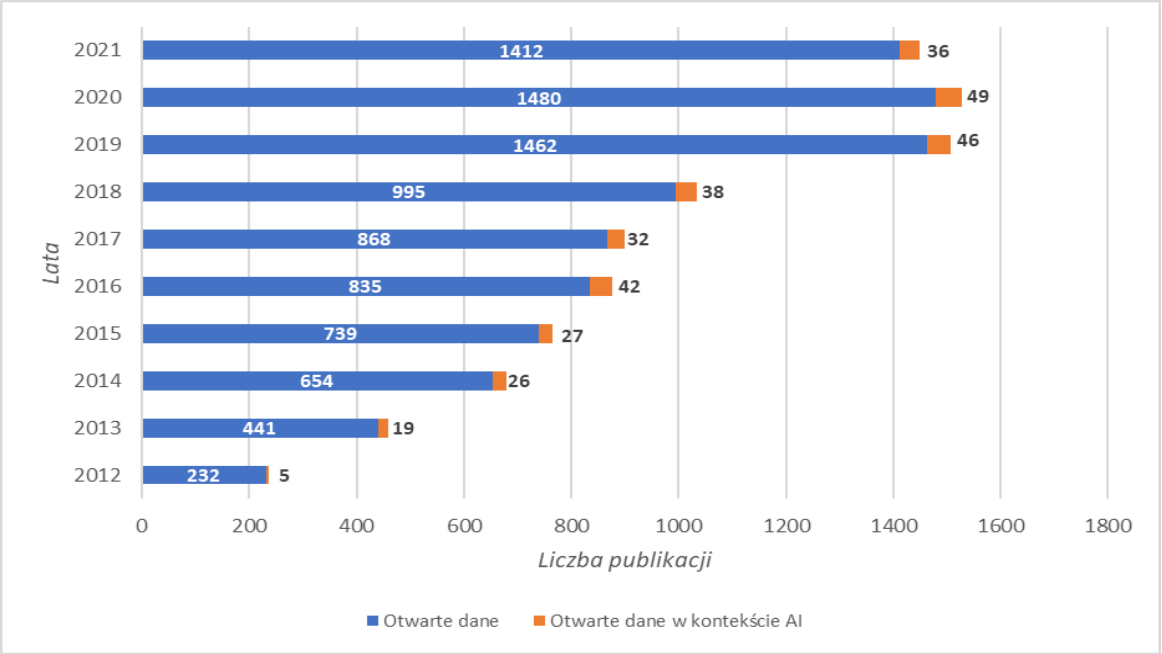
Kontekst badawczy

W celu identyfikacji obecnego stanu wiedzy i głównych kierunków badań w zakresie otwartych danych, z uwzględnieniem ich roli w rozwoju sztucznej inteligencji, przeprowadzono systematyczny przegląd literatury. Uwzględniono w nim trzy kluczowe obszary:

1. wyodrębnienie kluczowych baz danych i zbioru publikacji,
2. ich selekcję i opracowanie pełnej bazy publikacji,
3. ich analizę bibliometryczną i analizę treści.

W pierwszym etapie dokonano doboru baz danych, przyjmując jako kryterium ich pełnotekstowość, a ich analiza pozwoliła na wyodrębnienie ponad 9143 publikacji dotyczących otwartych danych, w tym 345 publikacji dotyczących otwartych danych, w kontekście AI, w latach 2011 – 2021 (stan na dzień: 16.09.2022 r.).

Wykres 1. *Liczba publikacji dotyczących otwartych danych oraz otwartych danych w kontekście rozwoju AI w latach 2011-2021*



Źródło: badania własne

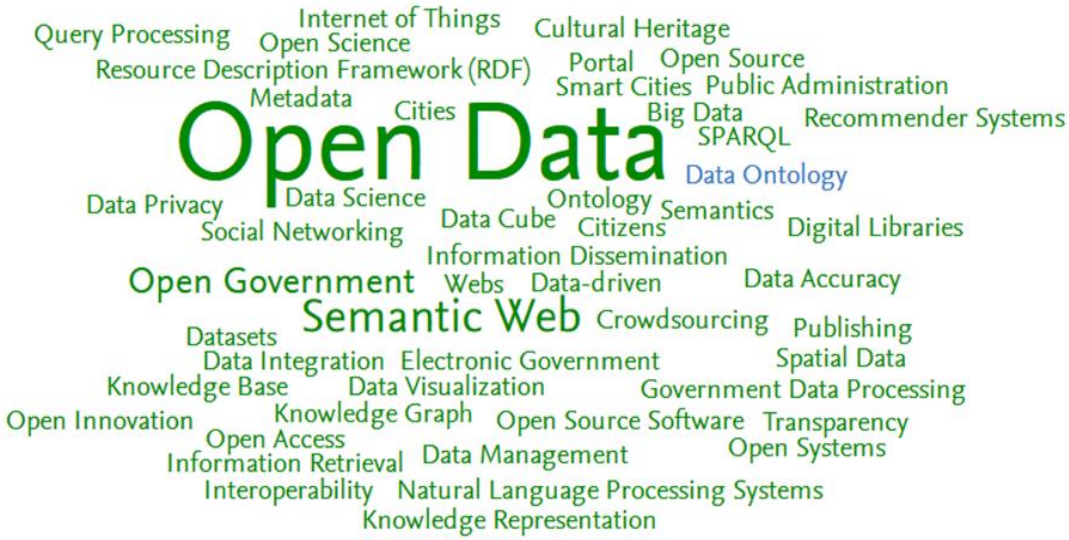
Analiza zarówno ilości publikacji dotyczących otwartych danych, jak i ilości tych publikacji, które odnoszą się bezpośrednio do problematyki otwartych danych na rzecz AI, wskazuje na ich znaczący wzrost w ostatnich latach. Wydaje się to w pełni uzasadnione obserwując z jednej strony rosnący poziom implementacji rozwiązań opartych na sztucznej inteligencji w różnych sferach naszego życia, jak i towarzyszący temu, z drugiej strony, rosnący popyt na dane. Prowadzi to do konieczności badań, ze

względu na identyfikację kolejnych luk naukowo-badawczych i ewolucję opisanych wcześniej procesów, zjawisk i zależności oraz popularyzację i komercjalizację uzyskanych wyników. Niewielki spadek liczby publikacji w 2021 roku wydaje się być efektem nie tyle spadku zainteresowania problematyką otwartych danych, a raczej efektem pandemii COVID-19 i przeniesienia w tym obszarze zainteresowania badaczy z wielu dziedzin i dyscyplin naukowych na problematykę związaną z jej przebiegiem, konsekwencjami, jak kształtowaniem się świata ery post-covid.

Wykaz dyscyplin podejmujących problematykę otwartych danych jest bardzo szeroki. Niewątpliwie dominują publikacje z zakresu informatyki (37,5%). Wśród pozostałych dyscyplin istotny udział odnotowują nauki społeczne (10,4%), nauki inżynieryjne (10,2%), matematyka (9,6%), a także biznes i zarządzanie z uwzględnieniem procesów podejmowania decyzji (7,8%). W przypadku otwartych danych na rzecz rozwoju sztucznej inteligencji, widać podobnie, jak dla publikacji z zakresu otwartych danych, dominującą pozycję opracowań z zakresu informatyki (49,2% wszystkich publikacji) i matematyki (17,5%). W dalszej kolejności pojawiają się nauki inżynieryjne (10,2%), opracowania z zakresu biznesu i zarządzania, z uwzględnieniem procesów podejmowania decyzji (7,6%) oraz nauki społeczne (5,2%).

Otrzymaną bazę publikacji poddano analizie z wykorzystaniem technik bibliometrycznych. Wykonano analizę frekwencyjności słów kluczowych oraz problemów badawczych. Wyselekcjonowane słowa kluczowe związane z analizowanymi frazami poddano analizie ilościowej. Jednocześnie wizualizację frekwencyjności przedstawiono w postaci *„chmury słów”*, gdzie częstotliwość występowania została odzwierciedlona w wielkości i grubości czcionki (Ilustracja 1 i Ilustracja 2).

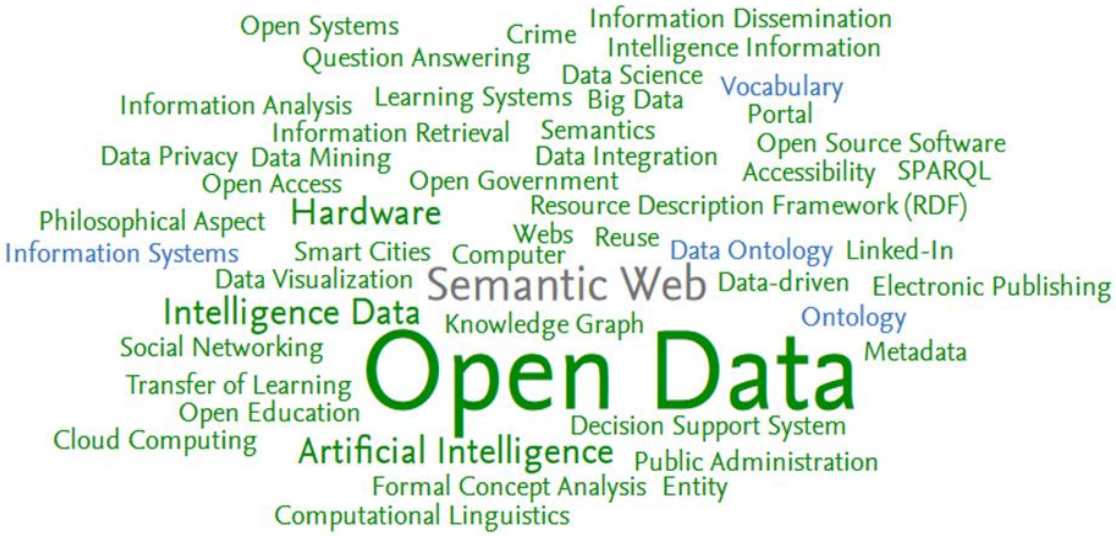
Ilustracja 1. *Analiza słów kluczowych dla frazy: otwarte dane (2011-2021)*



Źródło: badania własne

Najczęściej pojawiające się słowa kluczowe to: *Open Science* (wzrost liczby publikacji w analizowanym okresie = +3250%), *Smart Cities* (wzrost liczby publikacji w analizowanym okresie = +1160%), *Open Government* (wzrost liczby publikacji w analizowanym okresie = +587,57%) oraz *Metadata* (wzrost liczby publikacji w analizowanym okresie = +281,8%).

Ilustracja 2. *Analiza słów kluczowych dla frazy: otwarte dane na rzecz AI (2011-2021)*

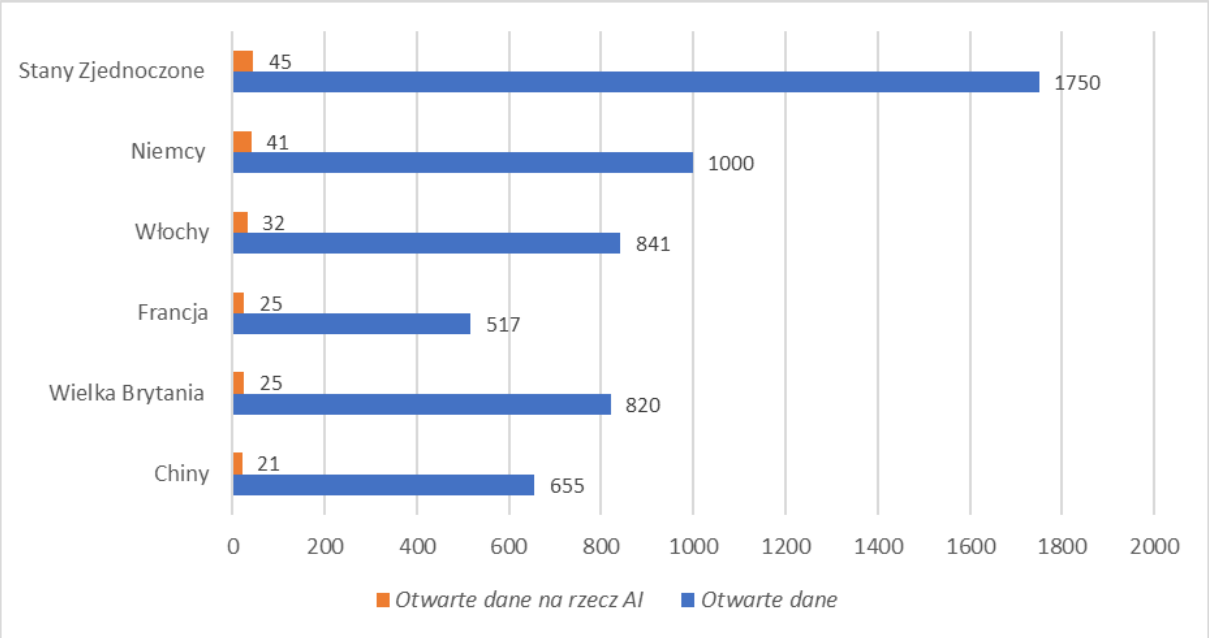


Źródło: badania własne

Najczęściej pojawiające się słowa kluczowe to oczywiście: *Artificial Intelligence* (wzrost liczby publikacji w analizowanym okresie = +600%), a także *Open Government* (wzrost liczby publikacji w analizowanym okresie = +500%) i *Intelligence data* (wzrost liczby publikacji w analizowanym okresie = +200%).

I chociaż problematyka otwartych danych wzbudza zainteresowanie badaczy w zasadzie na całym świecie, to analiza liczby publikacji z uwzględnieniem poszczególnych krajów wykazała, że najaktywniejsze w tym zakresie, zarówno dla problematyki otwartych danych, jak i otwartych danych na rzecz AI, są: Stany Zjednoczone, Niemcy, Włochy, Francja, Wielka Brytania i Chiny. Polska znalazła się w tej klasyfikacji na 27. pozycji wśród wyodrębnionych 144 krajów i regionów w zakresie problematyki otwartych danych oraz na 30. pozycji wśród 62 krajów i regionów zajmujących się problematyką otwartych danych na rzecz AI.

Wykres 2. *Najbardziej aktywne kraje według liczby publikacji dla problematyki otwartych danych i otwartych danych na rzecz AI za lata 2011-2021*



Źródło: badania własne

Wielowątkowość i złożoność zagadnień poruszanych w publikacjach z zakresu otwartych danych, jak i otwartych danych na rzecz AI, wymusza konieczność kompleksowego spojrzenia na te zagadnienia. Co więcej, procesy rynkowe w tym obszarze przebiegają dużo szybciej niż nauka jest w stanie je opisać i wyjaśnić ich przebieg. W efekcie wiele aspektów dotyczących otwartych danych w ujęciu społecznym czy biznesowym jest wciąż mało rozpoznana, co wymusza konieczność systematycznych i wnikliwych badań oraz analiz w tym zakresie.

Metodyka badań

Przeprowadzony przegląd literatury dotyczącej otwartych danych wskazuje na intensywny wzrost zainteresowania tą tematyką, co wydaje się uzasadnione biorąc pod uwagę dynamikę rozwoju technologii cyfrowych, w tym zwłaszcza sztucznej inteligencji oraz ich wpływ na rozwój społeczno-gospodarczy poszczególnych krajów i regionów. Dlatego też uznano za istotne określenie czynników determinujących decyzje przedsiębiorstw w zakresie wykorzystania dostępnych danych, w tym na rzecz rozwoju sztucznej inteligencji. Uwzględniono zarówno kwestie dotyczące świadomości i wiedzy, co do możliwości i form dostępu do danych publicznych, jak i zakresu ich praktycznej implementacji w działalności biznesowej. Zwrócono przy tym uwagę na analizę korzyści i kosztów, jakie przedsiębiorstwa dostrzegają w tym zakresie. Uznano to za istotne dla stworzenia szczegółowych ram i wymiarów metodologicznych niezbędnych do dalszych badań i pogłębionych analiz nad wykorzystaniem otwartych danych w strategii AI w Polsce.

Badanie miało charakter wstępny i było zorientowane na identyfikację postaw wobec otwartych danych w aktywności rynkowej przedsiębiorstw. Badania zrealizowano metodą wywiadów online (CAWI) w okresie w okresie październik 2021 r.-marzec 2022 r. Narzędziem badawczym był ustrukturyzowany kwestionariusz badawczy, składający się z 16 pytań, podzielonych na trzy bloki. W pierwszej części znalazły się pytania dotyczące zakresu korzystania z otwartych danych, ze szczególnym uwzględnieniem portalu Otwarte Dane. Druga część koncentrowała się na identyfikacji ograniczeń, barier i wyzwań, z którymi borykają się obecnie przedsiębiorstwa w analizowanym zakresie, z punktu widzenia rozwoju sztucznej inteligencji, zaś ostatnia odnosiła się bezpośrednio do cech respondentów (*województwo, wielkość firmy, sektor IT w jakim firma działa oraz główne źródło pochodzenia kapitału*).

Próba badawcza był niewielka, ze względu na charakter wstępny prowadzonych badań, co nie pozwala uogólniać wynikających z nich wniosków. Mają one charakter informacyjny i stanowią punkt wyjścia do dalszych pogłębionych studiów i analiz. Prezentowane w raporcie zagadnienia wpisują się w trwającą debatę nad procesem transformacji cyfrowej i szerokim spektrum jej następstw, w tym z punktu widzenia dalszych kierunków i dynamiki rozwoju sztucznej inteligencji i jej znaczenia społeczno-gospodarczego. Omówione zagadnienia nie wyczerpują podjętej problematyki, zarówno ze względu na ogromną różnorodność zagadnień, z jakimi mamy obecnie do czynienia, jak i dynamikę zachodzących w tym obszarze zmian. Dlatego będą one systematycznie powtarzane. Niewątpliwie konieczne jest zwiększenie próby i jej większe zróżnicowanie z punktu widzenia przyjętych zmiennych, w szczególności deskryptywnych i behawioralnych. Pozwoli to zarówno na nadbudowę zidentyfikowanych czynników determinujących decyzje o wykorzystaniu otwartych danych, w tym na rzecz rozwoju sztucznej

inteligencji, identyfikację zależności między nimi, jak i pozwoli ustalić istotne różnice między analizowanymi sektorami.

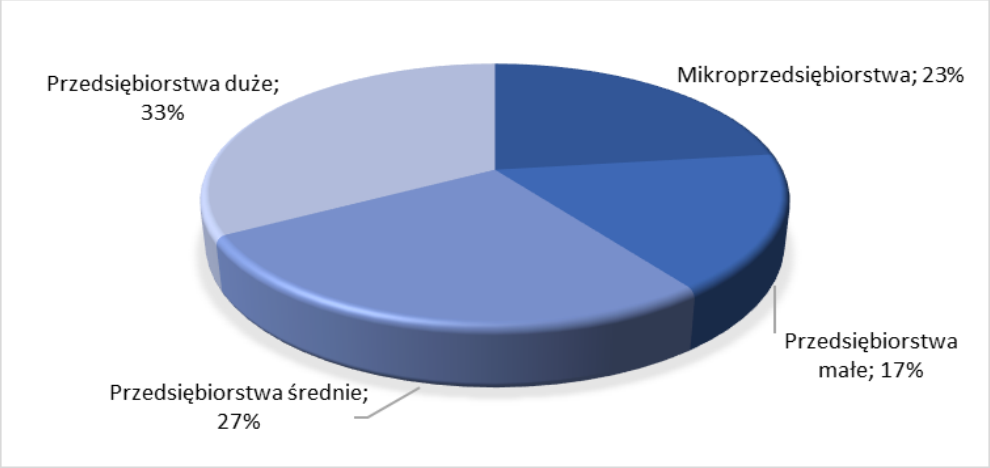
Raport powstał we współpracy firmy **Deviniti** i **Centrum Inteligentnych Technologii Wydziału Zarządzania Uniwersytetu Łódzkiego** w ramach prac Podgrupy ds. Badań, Innowacyjności i Wdrożeń Grupy Roboczej ds. Sztucznej Inteligencji.

Charakterystyka zbiorowości

Badanie zostało oparte na próbie 30 polskich przedsiębiorstw, które koncentrują się na opracowywaniu i wytwarzaniu nowych technologii lub usług dla swoich klientów. Dobór próby miał charakter celowy. Uznano bowiem, że ta grupa przedsiębiorstw ma wiedzę, umiejętności i kompetencje w analizowanym zakresie, mogąc stać się jednocześnie katalizatorem dalszych zmian.

W badaniu największy udział miały przedsiębiorstwa duże, które stanowiły nieco ponad 33% badanych oraz średnie (27%). Prawie ¼ stanowiły te należące do sektora mikroprzedsiębiorstw, a niespełna 17% stanowiły przedsiębiorstwa zaliczane do sektora małych. Znaczna część podmiotów biorących udział w badaniu miała swoją główną siedzibę zlokalizowaną w jednym z dwóch województw: mazowieckim (43%) lub dolnośląskim (37%).

Wykres 3. *Charakterystyka zbiorowości według deklarowanej wielkości firmy*

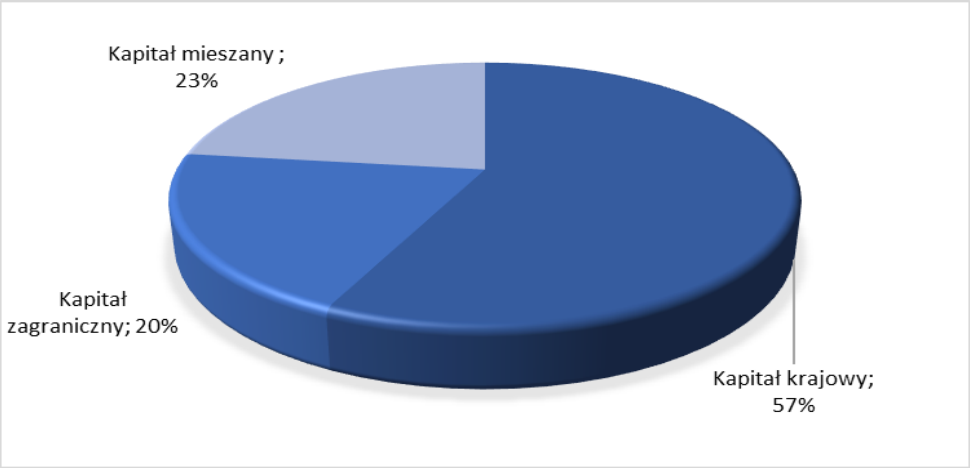


Źródło: badania własne

Prawie 47% przedsiębiorstw biorących udział w badaniu należało do jednego z trzech sektorów: AdTech (16,7%), Fintech (16,7%) oraz EduTech (13,3%). Wśród badanych znalazły się także firmy reprezentujące sektory takie, jak: BioTech, HealthTech i Medtech, e-Commerce oraz telekomunikacja.

W badanej grupie dominowały przedsiębiorstwa, których kapitał pochodził ze źródeł krajowych, stanowiąc 57% respondentów. Co czwarty badany podmiot deklarował, że kluczowym źródłem pochodzenia kapitału są źródła zagraniczne, zaś dla 23% był to kapitał mieszany.

Wykres 4. *Charakterystyka zbiorowości według źródeł pochodzenia kapitału firmy*



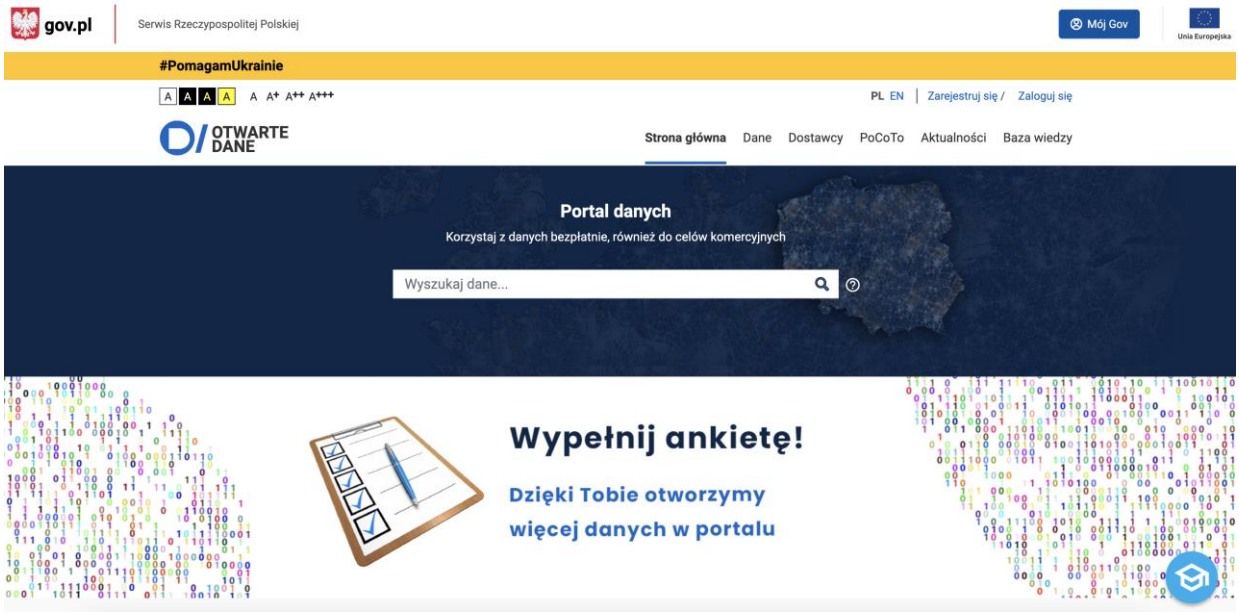
Źródło: badania własne

Przedsiębiorstwa biorące udział w badaniu deklarowały swoja zaawansowaną wiedzę w zakresie wykorzystania różnorodnych technologii cyfrowych, w szczególności tych zaliczanych do sztucznej inteligencji (np. Machine Learning, Deep Learning, Natural Language Processing).

Portal Otwarte Dane – dane.gov.pl

Portal Otwarte Dane jest oficjalnym serwisem Rzeczpospolitej Polskiej prezentującym wiarygodne i stale aktualizowane dane, pochodzące głównie z podmiotów administracji publicznej.

Ilustracja 3. *Strona główna portalu Otwarte Dane*



Źródło: https://dane.gov.pl/ [dostęp: 11.08.2022].

Charakterystyka wolumetryczna źródeł danych w serwisie to ponad 238 dostawców, 28 680 danych i 456 API (stan na 2022.11.15). Otwartość portalu przejawia się nie tylko w dostępnych danych, ale również w dostępie do kodów źródłowych serwisu, jak również, do pełnej dokumentacji API, w tym w formie interaktywnej. Serwis wskazuje także w formie przykładów na możliwości ponownego wykorzystania danych.

Jeśli chodzi o funkcjonalność, to jest ona standardowa dla tego typu serwisów. Mamy zatem dostęp do wyszukiwarki pełnotekstowej, a także katalogu z odpowiednio zaprojektowanymi filtrami. W większości przypadków istnieje możliwość podglądu danych (w tym wizualizacji graficznych) przed ich pobraniem.

Portal posiada również niezbędny zestaw narzędzi informacyjnych od biuletynu informacyjnego, bazy wiedzy po aktualności, co powinno mieć wpływ na propagowanie wiedzy w tym zakresie i podnoszenie związanych z tym umiejętności, zwłaszcza w odniesieniu do przedsiębiorstw. Ciekawym narzędziem w tym zakresie jest zestawienie produktów, aplikacji, systemów, stron internetowych wykorzystujących dane z portalu Dane.gov.pl.

Co drugie przedsiębiorstwo biorące udział w badaniu znało portal Otwarte Dane. Fakt ten nie okazał się jednak jednoznaczny z ich decyzją o korzystaniu z udostępnianych tam danych publicznych. To potwierdziło zaledwie 20% badanych przedsiębiorstw. Analiza wyników badania według przyjętych cech respondentów wskazała, że w grupie tej znalazły się duże przedsiębiorstwa, działające w sektorze FinTech lub EduTech, które deklarowały wykorzystanie co najmniej jednej techniki lub metody AI. Pochodzenie kapitału oraz województwo, w którym firma ma swoją główną siedzibę miały znaczenie marginalne.

Wśród najczęstszych przyczyn dla których przedsiębiorstwa nie korzystały z danych udostępnianych na portalu Otwarte Dane na potrzeby realizowanych technik, algorytmów i metod AI przede wszystkim wskazywano na fakt, że *„duża ich część jest prezentowana po nadmiernym przetworzeniu i zagregowaniu, co niejednokrotnie sprawia, że są one mało przydatne do rynkowego zastosowania w przypadku konkretnych problemów”.* W opinii badanych firm chociaż *„portal prezentuje dużo danych są one niejednokrotnie nieużyteczne z punktu widzenia samego rozwoju i potrzeb związanych z AI”,* ze względu na fakt, że wiele zbiorów ma charakter *„wyrywkowy”, „często lokalny”* i *„nie dotyczy całego kraju”.* Respondenci uznali także, że portal *„nie daje możliwości pobierania przyrostów, obrazujących zachodzące zmiany, w tym w czasie rzeczywistym, a jedynie cały zakres informacji*”, a udostępniane zbiory danych są *„często zbyt małe, jak na potrzeby aktywności związanej z rozwojem AI”.* Jako ograniczenie wskazano także *„brak dostępu do danych źródłowych w przypadku części danych”,* co może wynikać z zasad udostępniania tych danych.

Znaczenie danych w tworzeniu rozwiązań opartych o sztuczną inteligencję, w szczególności algorytmy nauczania maszynowego, jest nie do przecenienia w obecnych czasach, w szczególności w ostatniej dekadzie, kiedy zostały spopularyzowane metody nauczania głębokiego (deep learning) przy jednoczesnym zwiększeniu dostępności zasobów sprzętowych takich, jak karty graficzne NVIDIA z platformą CUDA, które pozwoliły na osiąganie doskonałych wyników w dziedzinie rozpoznawania obrazów, przetwarzania tekstu naturalnego czy analizy sygnałów wideo, bez potrzeby angażowania zasobów dostępnych jedynie dla ośrodków akademickich, organizacji rządowych czy dużych firm. Obecnie nawet małe firmy mogą pozwolić sobie na wdrażanie innowacyjnych rozwiązań względnie niewielkim kosztem.

Wszystkie te nowe możliwości na nic jednak się zdadzą, jeżeli organizacja nie dysponuje danymi. Oczywiście, bardzo często kluczowe jest posiadanie czy też przygotowanie danych charakterystycznych właśnie dla danej organizacji, co wiąże się z potrzebą włożenia pewnego wysiłku po jej stronie. W wielu przypadkach, zwłaszcza na początku realizacji projektów badawczych, gdy konieczne jest stworzenie choćby wczesnego prototypu i wykazanie, czy postawiona hipoteza projektowa ma rację bytu, otwarte zbiory danych pozwalają na szybką

weryfikację pomysłu. Co więcej, nawet w projektach dojrzałych, dysponujących własnym, autorskim zestawem danych, otwarte zbiory danych mogą być nieocenione w sytuacji, gdy pewna część rozwiązania wymaga stworzenia modelu opartego o dane, których organizacja po prostu nie ma. O ile w przypadku analizy obrazów medycznych skany RTG, wyniki badań CT czy MRI mają powszechnie, międzynarodowo stosowane formaty, o tyle w przypadku przetwarzania tekstu naturalnego, pojawia się oczywista bariera językowa i najlepiej opracowana baza (korpus) języka angielskiego nie zastąpi nam bazy języka polskiego. Nie dziwi więc obecność tego elementu w najbardziej brakujących rodzajach danych przedstawionych w badaniu.

Nie pozostaje się również nie zgodzić z wykazanym przez respondentów brakiem powiązań między cechami a etykietami. Koniec końców to właśnie uczenie z nadzorem (supervised learning) stanowi jedną z najistotniejszych gałęzi nauczania maszynowego, a w szczególności głębokiego. Przygotowanie powiązań między cechami a etykietami to nierzadko żmudna praca wykwalifikowanego specjalisty (np. w przypadku oznaczania badań medycznych na różnego rodzaju skanach/obrazach często konieczne jest po prostu przeprowadzenie diagnozy na podstawie tychże obrazów). Często osoby czy organizacje po wytworzeniu takiej bazy nie chcą się nimi dzielić, wychodząc z założenia, że poniosły one spore koszty i stąd też chcą zostawić tę wartość u siebie. Niewątpliwie byłoby korzystnym, gdyby zaczęto w takich sytuacjach stosować się choćby do obyczajów znanych branży IT w kontekście otwartego oprogramowania (open source), gdzie nie tylko niezależne zespoły programistyczne, ale i duże firmy potrafią dzielić się efektami swojej pracy z innymi programistami na całym świecie kompletnie za darmo, budując swój model biznesowy i model korzyści wokół oprogramowania, a nie na zasadzie czerpania bezpośrednich przychodów z dostępu do programów (licencjonowania).

dr inż. Krzysztof Rychlicki-Kicior
Chief Scientific Officer
Makimo sp. z o.o.

Biorąc pod uwagę dalszy rozwój portalu, warto w opinii respondentów, zwrócić uwagę na potrzeby tych przedsiębiorstw, które już działają w obszarach związanych ze sztuczną inteligencją lub będą chciały związać z nią swój model biznesowy w przyszłości. Dotyczy to przede wszystkim:

- poszerzenia spektrum udostępnianych danych (np. *„szczegółowe dane demograficzne”, „źródłowe dane pochodzące z Głównego Urzędu Statystycznego”, „informacje o ruchu miejskim”, „dane dotyczące poziomu produkcji w różnych sektorach”*)
- wzrostu jakości udostępnianych danych z punktu widzenia potrzeb AI (np. *„dane prezentowane w czasie rzeczywistym”, „możliwość pobierania jedynie kolejnych przyrostów obrazujących zmianę”, „dane multimedialne, w tym tekstowe, głosowe przygotowane do uczenia maszynowego*)

- kwestii prezentacji zbiorów danych na stronie portalu (np. *„ocena ekspercka poziomu przydatności konkretnych zbiorów z punktu widzenia AI”, „zakładki dedykowane na potrzeby AI”, „moderacja ekspercka”*).

Polska polityka dotycząca dostępności otwartych danych, w tym działania realizowane za pośrednictwem portalu Otwarte Dane, niewątpliwie zasługują na uznanie, co potwierdza chociażby wysoka pozycja Polski w raporcie *Open Data Maturity 2021* czy w *Indeksie Gospodarki Cyfrowej i Społeczeństwa Cyfrowego DESI 2022 (Polska jest powyżej średniej dla wszystkich państw UE).* Co więcej, realizowany obecnie program *„Otwarte dane plus”* ma doprowadzić do znacznego zwiększenia ilości i poprawy jakości udostępnianych danych publicznych przyczyniając się także do intensyfikacji ich ponownego wykorzystania. Przedstawione w raporcie opinie respondentów dotyczące portalu Otwarte Dane należy zatem traktować jako sugestie co do możliwości dalszego jego rozwoju na rzecz zwiększenia dynamiki AI w Polsce i multiplikacji wynikających z tego korzyści.

Należy postulować dalszy rozwój centralnego punktu dostępu do danych (dane.gov.pl). W konsekwencji poszerzany powinien być także sam katalog zbiorów dostępnych przez ten punkt. Z kolei w przypadkach, w których nie będzie to możliwe, jak najwięcej zbiorów powinno być udostępnianych bezwnioskowo oraz w miarę potrzeby, w czasie rzeczywistym (np. przez API).

Z kolei tworzenie systemu współdzielenia danych pochodzących z różnych źródeł (w tym zwłaszcza ze źródeł prywatnych) wymaga przede wszystkim zapewnienia warunków dla ich bezpiecznej wymiany. Należy przez to rozumieć bezpieczeństwo dla samych danych (np. ochronę danych osobowych przed wyciekiem lub bezprawnym wykorzystaniem). Nie mniej ważne jest jednak bezpieczeństwo regulacyjne dla podmiotów zaangażowanych w taką działalność. Administratorzy systemów współdzielenia powinni wiedzieć, że stosując określone wymagania nie będą narażeni na ryzyka prawne (zwłaszcza ze strony regulatorów takich jak PUODO, UOKiK, UKE i in.). Można to osiągnąć, przykładowo, poprzez zaangażowanie odpowiednich władz (w tym w szczególności ww. regulatorów) w określenie precyzyjnych i praktycznych wymogów dla takiej działalności (np. zasad prawidłowej anonimizacji oraz jej możliwych obszarów czy zdefiniowanie innych kryteriów przetwarzania danych, których spełnienie będzie chronić zaangażowane podmioty przed ryzykiem sankcji).

Oczywiście konieczne jest także ustalenie statusu danych zebranych w takim systemie oraz ścisłe określenie praw i obowiązków podmiotu, który nim zarządza. Istotne jest również zapewnienie odpowiedniego, zestandaryzowanego środowiska informatycznego (w tym określenie zasad interoperacyjności systemu z systemami zewnętrznymi czy uzgodnienie właściwych standardów/formatów wymiany danych, dzięki którym system będzie mógł funkcjonować jako całość).

Dodatkowo, zasada otwartości powinna być realizowana w perspektywie między- oraz ponadnarodowej. Oznacza to nie tylko zapewnienie dostępu do danych dla partnerów zagranicznych (przy założeniu wzajemności z ich strony), ale i dalszy udział we wspólnych projektach realizujących ideę wspólnych przestrzeni danych.

Piotr Zawadzki
Radca prawny i rzecznik patentowy w kancelarii Bird&Bird

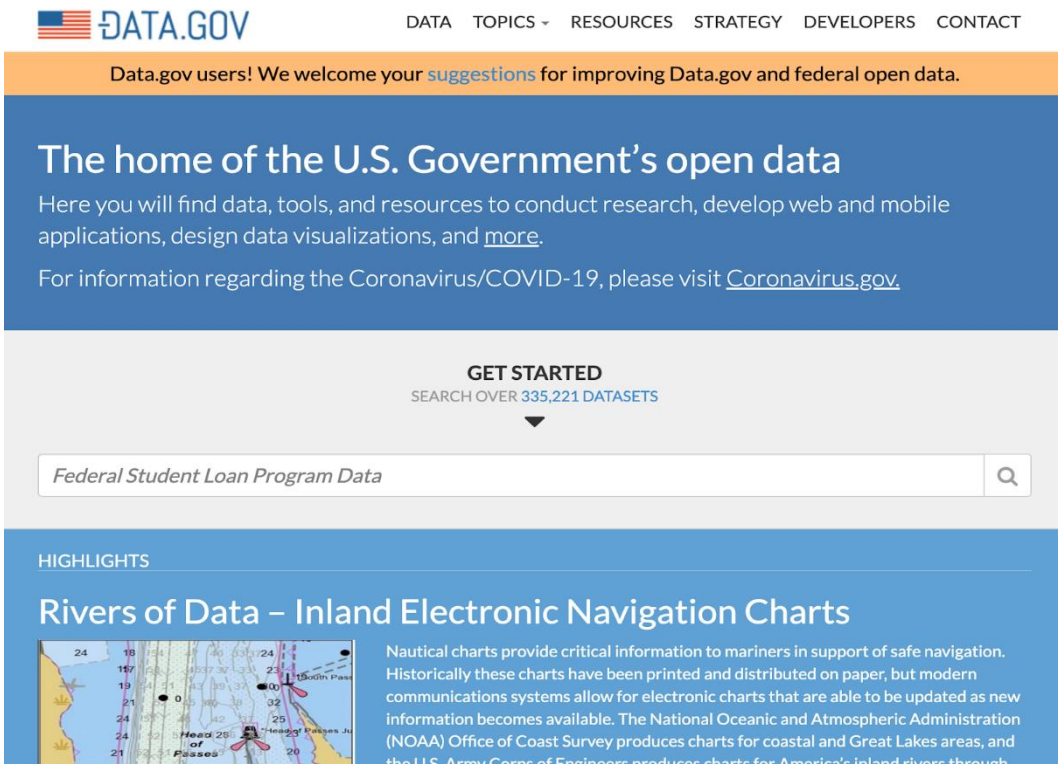
Wybrane przykładowe zagraniczne portale dostarczające otwartych danych

Warto zwrócić jednocześnie uwagę na dostępność zagranicznych portali udostępniających otwarte dane, które mogą mieć znaczenie dla polskiego biznesu, w tym w kontekście dalszego rozwoju sztucznej inteligencji.

Portal https://data.gov/

Nasz przegląd zagranicznych portali otwartych danych rozpoczniemy od jednego z pionierów na świecie tj. https://data.gov/, czyli otwarte dane rządu USA i poszczególnych stanów.

Ilustracja 4. *Strona główna portalu data.gov*



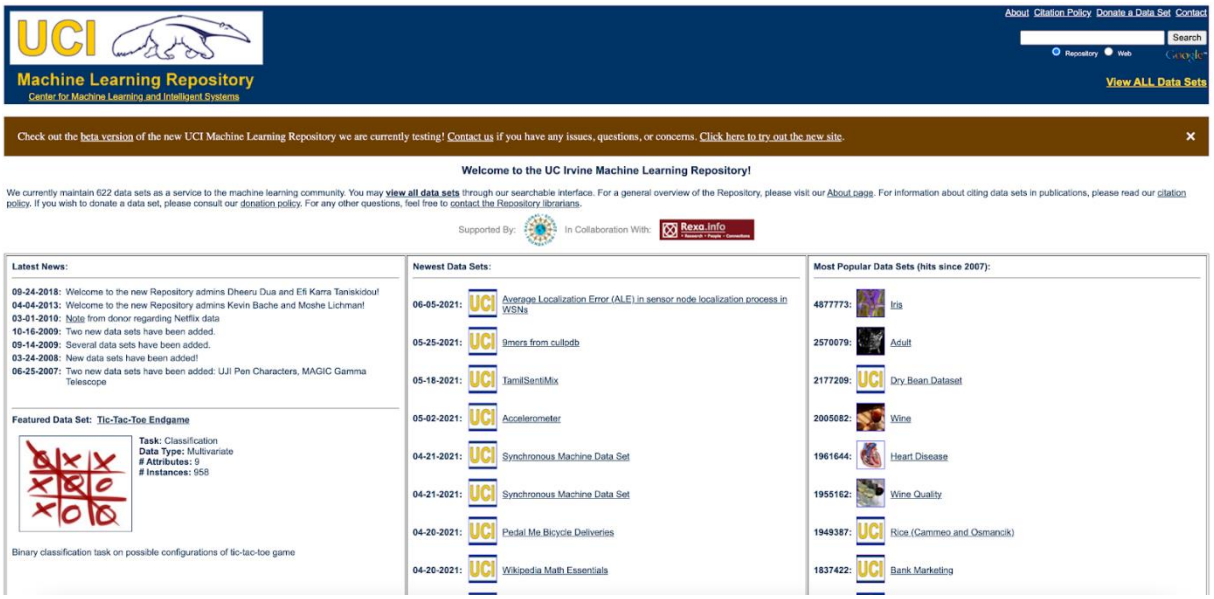
Źródło: https://data.gov/ [dostęp: 11.08.2022].

Portal ten powstał kilka miesięcy po wyborze Baracka Obamy na prezydenta USA, który zaczął swoją prezydenturę od podpisania trzech memorandów prezydenckich. Dwa z nich dotyczą otwartego rządu opartego zwłaszcza na otwartych danych. Skutkiem tych inicjalnych działań było wiele inicjatyw w obszarze "open" m.in. witryna i system https://data.gov/. Portal zawiera ponad 335 000 zestawów danych. Amerykański portal posiada bardzo prostą funkcjonalność tj. wyszukiwarka pełnotekstowa, katalog z rozbudowanym filtrowaniem i przeglądanie szczegółów poszczególnych zestawów danych. Zatem podobną funkcjonalność jaką oferuję polski portal otwartych danych czy inne europejskie odpowiedniki. Portal https://data.gov/ oferuje jednak znacznie więcej danych często w specjalistycznych formatach, np. KML, chemical/x-mdl-sdfile, geoJSON, GML itp. Posiada duże ilości danych o niewielkim stopniu przetworzenia w formatach branżowych, co zwiększa ich przydatność np. do uczenia maszynowego. Mimo to w opisie danych niestety brakuje takiej informacji, nie ma jej także w rozszerzonym opisie w filtrach i katalogach portalu.

Portal https://archive.ics.uci.edu/

Archiwum zbiorów danych Uniwersytetu Kalifornijskiego w Irvine gromadzi otwarte dane, których głównym zastosowaniem jest uczenie maszynowe. Zbiór ten nie jest duży – ponad 600 zbiorów, jednak są one bardzo starannie skategoryzowane pod względem AI.

Ilustracja 5. *Strona główna archiwum zbiorów danych uniwersytetu UCI.*



Źródło: *https://archive.ics.uci.edu/* [dostęp: 11.08.2022].

W filtrach katalogu zbioru mamy takie parametry jak:

- domyślne zadanie np. klasyfikacja, regresja, klastrowanie;
- typy atrybutów: kategoryczne, numeryczne;
- liczbę atrybutów oraz liczbę obiektów w zestawie;
- typ danych np. tekstowe, szeregi czasowe, sekwencje itp.;

Ponadto otwarte dane pochodzą nie tylko ze źródeł rządowych, ale także od firm komercyjnych, np. Amazon, MS NBC, CNN, Twitter i wielu innych. Mimo, niewielkiej liczby zbiorów i ich statycznego charakteru, dane zgromadzone w archiwum mają ogromny potencjał edukacyjny. Trudno wyobrazić sobie lekcję czy szkolenie z uczenia maszynowego bez wizyty na portalu ICM.

Portal DataPortals.org

DataPortals.org nie jest archiwum gromadzącym otwarte dane, tylko listą portali otwartych danych na całym świecie. Portal prowadzony jest przez Fundację Open Knowledge, która skupia czołowych ekspertów w dziedzinie otwartych danych z całego świata. Pierwsza wersja portalu została uruchomiona w 2011 roku w Berlinie. Lista obejmuje blisko 600 portali z otwartymi danymi, starannie wykonanymi opisami i cechami (np. język, kraj, status itp.).

Ilustracja 6. *Strona główna DataPortals.org*

Źródło: https://dataportals.org/ [dostęp: 11.08.2022].

Portal https://data.europa.eu/

Oficjalny portal otwartych danych Unii Europejskiej agregujący katalogi (portale krajowe), jak również zbiory danych z poszczególnych portali. Natomiast samo pobieranie, po wyszukaniu odpowiedniego zbioru, realizowane jest z krajowego/lokalnego źródła. Zaletą tego rozwiązania jest możliwość

korzystania np. z polskiego interfejsu użytkownika i przeglądania/wyszukiwania danych z dowolnego kraju Unii Europejskiej w wybranym języku.

Ilustracja 7. *Strona główna* https://data.europa.eu/

Źródło: https://data.europa.eu/ [dostęp: 11.08.2022].

Dane pochodzą z 36 krajów i są podzielone na 173 katalogi. W efekcie portal kategoryzuje imponujący zbiór zbiorów danych, tj. 1 437 988. Opisy skatalogowanych danych są automatycznie tłumaczone na wybrany język, co jest bardzo pomocne przy ich przeglądaniu manualnym. Warto tu zaznaczyć, że jednak pełnotekstowa wyszukiwarka portalu ma czasem problemy z automatycznie tłumaczonymi tekstami, w których tłumaczy nazwy własne i terminy informatyczne, zniekształcając jednocześnie wyniki wyszukiwania. Przykładowo, szukając danych związanych z energią jądrową, otrzymujemy na przykład: *pasza atomowa gminy Föritztal*, co jest źle przetłumaczonym opisem danych, czyli informacji przestrzennej gminy Föritztal w formacie *AtomFeed*.

Portale z otwartymi danymi poszczególnych krajów UE

Lp.	Kraj	Oficjalny portal z otwartymi danymi
1.	Austria	https://www.data.gv.at/
2.	Belgia	https://data.gov.be/
3.	Bułgaria	https://data.egov.bg/
4.	Chorwacja	https://data.gov.hr/

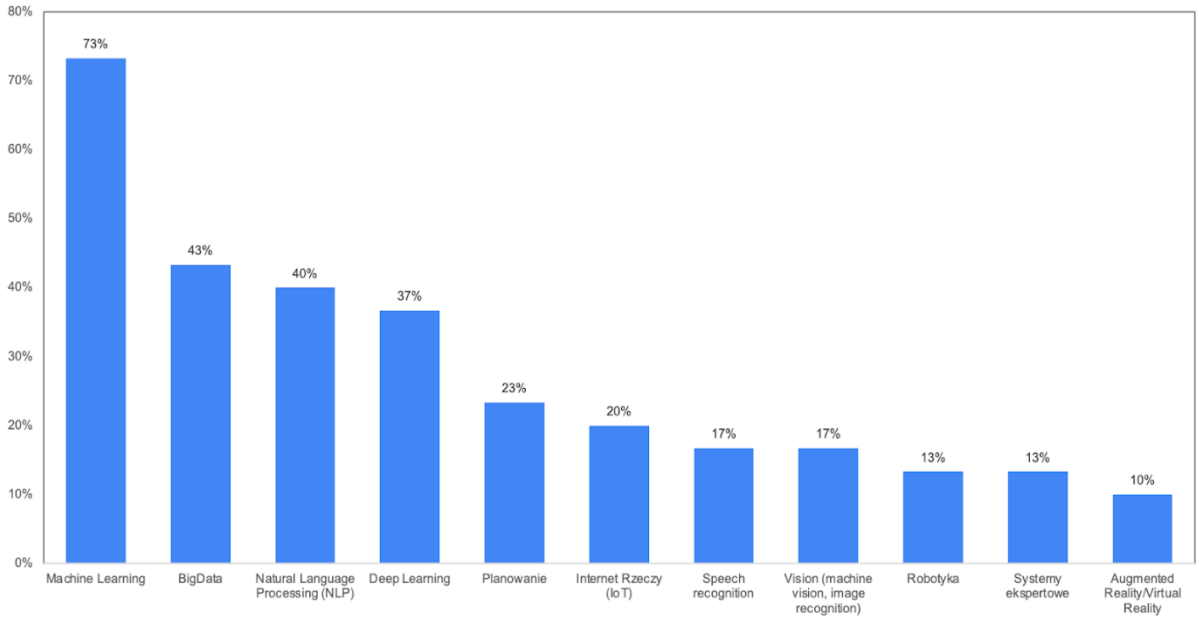
5.	Cypr	https://www.data.gov.cy/
6.	Czechy	https://data.gov.cz/
7.	Dania	https://www.opendata.dk/
8.	Estonia	https://opendata.riik.ee/
9.	Finlandia	https://www.avoindata.fi/
10.	Francja	https://www.data.gouv.fr/
11.	Grecja	https://www.data.gov.gr/
12.	Hiszpania	https://datos.gob.es/
13.	Holandia	https://data.overheid.nl/
14.	Irlandia	https://data.gov.ie/
15.	Litwa	https://data.gov.lt/
16.	Luksemburg	https://data.public.lu/
17.	Łotwa	https://data.gov.lv/
18.	Malta	https://open.data.gov.mt/
19.	Niemcy	https://www.govdata.de/
20.	Polska	https://dane.gov.pl/
21.	Portugalia	https://dados.gov.pt/
22.	Rumunia	https://data.gov.ro/
23.	Słowacja	https://data.gov.sk/
24.	Słowenia	https://podatki.gov.si/
25.	Szwecja	https://data.riksdagen.se/ https://www.dataportal.se/
26.	Węgry	b.d.
27.	Włochy	https://www.dati.gov.it/

Źródło: opracowanie własne

Potrzeby organizacji w obszarze otwartych danych

Firmy biorące udział w badaniu to firmy technologiczne o wysokim wskaźniku wykorzystania AI. Ponad 73% badanych wykorzystuje w swojej organizacji uczenie maszynowe, przetwarzanie języka naturalnego (40%) oraz deep learning (36,7%).

Wykres 5. *Kluczowe dziedziny, techniki, algorytmy wykorzystujące AI w organizacjach respondentów*



Źródło: badania własne

Wszystkie te wyżej wymienione techniki wymagają najczęściej tzw. danych etykietowanych (labeled data). Etykietowanie danych lub adnotacja danych jest częścią etapu wstępnego przetwarzania danych podczas tworzenia modeli uczenia maszynowego (ML). Wymaga identyfikacji surowych danych (obrazów, plików tekstowych, filmów), a następnie dodania jednej lub więcej etykiet do tych danych, w celu określenia ich kontekstu dla modeli, umożliwiając modelowi uczenia maszynowego dokonywanie dokładnych przewidywań. Zatem najczęściej dane tego typu można uzyskać poprzez „ręczne" cechowanie obiektów etykietami w surowych zestawach danych. Oczywiście, z tego powodu uzyskanie danych oznaczonych etykietami jest znacznie droższe niż uzyskanie „surowych" danych statystycznych.

Oczywiście istnieją techniki niewymagające oznaczonych danych np. klastrowanie (uczenie nienadzorowane) czy też uczenie przez wzmacnianie (reinforcement learning, RL). Jednak w biznesie wciąż prym wiodą algorytmy klasyfikacji (zwłaszcza tekstu), regresji oraz kategoryzacja obrazów czy też identyfikacja obiektów na obrazie np. w dziedzinie komputerowego widzenia. We wszystkich tych

algorytmach do osiągnięcia założonych celów zaimplementowanych metod niezbędne są dane etykietowane, jak również większa ingerencja nadzoru ludzkiego.

W tym obszarze respondenci wykazali największy niedobór publicznie dostępnych otwartych danych. Ponad 46% respondentów stwierdziło, że w ich branży, w Polsce brakuje danych przydatnych do uczenia maszynowego w postaci cechy-etykieta. Drugim istotnym rodzajem danych, których brakuje przedsiębiorcom w skutecznym wdrażaniu AI, okazały się ogólnodostępne korpusy i słowniki języka polskiego, wskazane przez 23,3% respondentów. Przedsiębiorcy wskazywali również na braki w polskich repozytoriach otwartych danych w kategoriach takich jak: zdjęcia, pliki dźwiękowe czy filmowe ze szczegółowymi opisami.

Należy zauważyć, że dane etykietowane (jakościowo i wolumetrycznie), a później zbudowane na nich efektywne modele uczenia maszynowego stanowią istotną przewagę konkurencyjną firmy. Tym samym powinny być one inwestycją własną firmy. Jednak to właśnie otwarte dane etykietowane są warunkiem koniecznym skutecznego podnoszenia kompetencji w zakresie sztucznej inteligencji polskich inżynierów. Niestety, repozytoria otwartych danych z Polski i innych krajów Unii Europejskiej nie posiadają etykietowanych zestawów obrazów, tekstów czy filmów. Nie udostępniają też bezpłatnych narzędzi do etykietowania (anotowania) udostępnianych przez siebie danych.

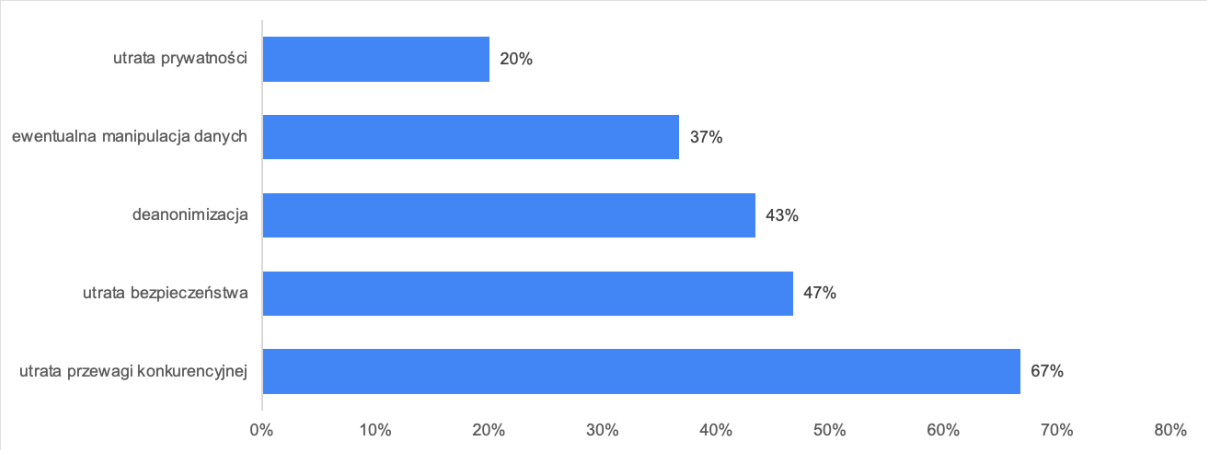
Gotowość udostępniania danych przez przedsiębiorców

Coraz więcej zagranicznych firm (głównie z obszaru e-commerce) udostępnia swoje dane na otwartej licencji. Dobrym przykładem jest Amazon, który nie tylko udostępnia ponad 350 otwartych zbiorów danych (w tym medycznych), gdzie sponsorem jest ich spółka zależna Amazon Web Services (AWS), ale także dane ze swojego flagowego sklepu (amazon.com), np. gotowe do uczenia maszynowego dane w postaci 130 milionów recenzji klientów (wraz z metadanymi/etykietami) zebrane przez sklep w latach 1995-2015. Dane te stanowią bogate źródło informacji dla badaczy akademickich m.in. w dziedzinie przetwarzania języka naturalnego (natural language processing, NLP), ekstrakcji informacji (information extraction) oraz uczenia maszynowego. Niestety głównym językiem recenzji jest język angielski, a w zbiorach danych nie ma polskich recenzji. Dlatego też, głównie z potrzeb edukacyjnych, od polskich przedsiębiorców wymagana jest coraz większa elastyczność w zakresie udostępniania własnych zbiorów danych na otwartej licencji, w sposób kontrolowany (bez danych osobowych i wrażliwych).

Niestety, 51,7% respondentów nie potrafiło jednoznacznie odpowiedzieć, czy ich organizacja jest w stanie nie tylko korzystać z otwartych danych, ale także nimi się dzielić. Zdecydowanie, *tak* odpowiedziało 27,6%, a *nie* tylko 20,7%.

Dla 67% badanych firm najważniejszą obawą związaną z otwarciem danych była utrata przewagi konkurencyjnej. Natomiast nieco ponad 47% respondentów wskazało na utratę bezpieczeństwa lub ryzyko deanonimizacji (43%), np. odtworzenie danych osobowych lub wrażliwych (sprzedaż, marża) z wcześniej zamaskowanych danych.

Wykres 6. *Główne obawy badanych przedsiębiorców w kontekście otwierania danych*



Źródło: badania własne

Należy zauważyć, że nawet przy obecnych zaawansowanych technikach automatycznej anonimizacji danych, koszt przygotowania danych do "otwarcia" będzie znaczący. Ponadto będzie on wymagał wysokiego poziomu kontroli manualnej, np. przez wyspecjalizowanych pracowników. Źle przeprowadzony proces "uwalniania" danych przez organizację nawet w szczytnym celu, jak edukacja czy podnoszenie konkurencyjności polskiego biznesu wobec firm zagranicznych, może skutkować realizacją obaw wymienionych przez respondentów powyżej i wystąpieniem szeregu zagrożeń prawnych, reputacyjnych lub utraty przewag konkurencyjnych.

Projektowanie, implementacja i wdrożenie mechanizmów wspierających proces przetwarzania danych jest dla organizacji wielkim wyzwaniem. Współcześnie praca z danymi, bez względu na źródło ich pochodzenia, format czy wolumen, należy do zadań nietrywialnych. Dane, będące w posiadaniu organizacji, stanowią często znaczącą część wartości jej biznesu. Jednocześnie z danymi nierozłącznie związane jest oprogramowanie, wykorzystane podczas ich zbierania, przetwarzania, przesyłania i prezentacji. Nie dziwi więc fakt braku pewności respondentów odnośnie stopnia zaangażowania ich organizacji w ustanowienie otwartego dostępu do ich zasobów.

Współdzielenie zbiorów danych, oprócz szeregu kwestii związanych z ich ochroną, jest również interesującym zagadnieniem z zakresu inżynierii oprogramowania. Udostępnienia danych nie

należy jedynie utożsamiać z przechowaniem pewnego zasobu w miejscu umożliwiającym swobodny do niego dostęp. Współdzielenie jest pojęciem szerszym. Aby doszło do otwarcia danych należy najpierw ustanowić ujednolicony proces ich zbierania oraz etykietowania. Dane powinny zostać poddane wstępnemu przetworzeniu w celu ich walidacji, oczyszczenia z błędów, uzupełnienia braków, normalizacji itd. Dopiero wtedy jesteśmy gotowi do procesu zapisu danych i wystawienia ich na zewnątrz. Na tym praca się nie kończy. Aby opublikowany zbiór danych miał wartość badawczą należy w czytelny sposób przedstawić metodykę uzyskania danych i szczegółowo opisać proces ich przetwarzania. Oprogramowanie przetwarzające dane (data processing pipeline) powinno być wersjonowane, ponieważ jego przyszła zmiana może spowodować wprowadzenie różnić między obserwacjami uzyskanymi w różnym czasie. Co więcej, dostęp do zbioru danych powinien być stale monitorowany w celu poznania sposobu wykorzystania udostępnionego zasobu. Sam zbiór powinien być regularnie aktualizowany i poddawany weryfikacji, aby upewnić się, że jest on nadal aktualny i przystaje do postawionej przed nim gamy problemów badawczych (data drift). Przejście wszystkich wymienionych kroków na drodze do udostępnienia danych jest kosztowne i wymaga dużo uwagi ze strony organizacji decydującej się na współdzielenie danych.

dr Maciej Adamiak
Chief Scientific Officer w ReasonField Lab

Oprócz dalszego otwierania danych, warto postulować zwiększanie aktywności sektora publicznego na polu tworzenia baz wysokiej jakości. Problemem systemów AI jest bowiem nie tylko brak dostępu do danych, ale także różnego rodzaju wady i inne ograniczenia dotyczące teoretycznie dostępnych zbiorów, a które mogą sprzyjać zachwianiu wiarygodności lub przydatności wyników ich zastosowania.

Z punktu widzenia podmiotu, który pragnie wykorzystać dane na potrzeby systemu AI wadą mogą być zarówno anomalie zbiorów (prowadzące do zjawisk określanych jako tzw. AI biases), jak i ograniczenia prawne związane ze statusem danych (np. ograniczenia dotyczące danych osobowych czy ochrona wynikająca z praw własności intelektualnej, w tym tzw. praw sui generis do baz danych), które czynią zbiory mniej wartościowymi.

Dane wykorzystywane na potrzeby systemów AI powinny natomiast odpowiadać określonym wymogom. Najczęściej wskazuje się w tym kontekście na konieczność spełniania przez zbiory tzw. wymogu 5 V, czyli Variety (różnorodność struktury i treści), Volume (objętość zbioru, możliwie jak największa), Veracity (wiarygodność i dokładność), Velocity (szybkość narastania zbioru), a także Value (faktyczna wartość). Tworząc bazy danych na potrzeby ich udostępnienia, stosowni decydenci (w tym w szczególności odpowiednie władze publiczne), powinni więc, już w fazie projektowania baz lub zasad dostępu do nich, dbać o spełnianie przez nie opisanych wyżej wymogów.

Krokiem w dobrą stronę byłaby także efektywna anonimizacja danych (zwłaszcza danych osobowych) przez ich publicznych dysponentów (przed ich udostępnieniem zainteresowanym) oraz tworzenie (i udostępnianie) baz danych już zanonimizowanych. Decydenci powinni także

w sposób bardziej świadomy stosować istniejące wyjątki od ochrony zapewnianej przez prawa własności intelektualnej, w tym zasady dozwolonego użytku publicznego i in. reguły pozwalające na wykorzystanie danych, co do zasady, chronionych.

Piotr Zawadzki
Radca prawny i rzecznik patentowy w kancelarii Bird&Bird

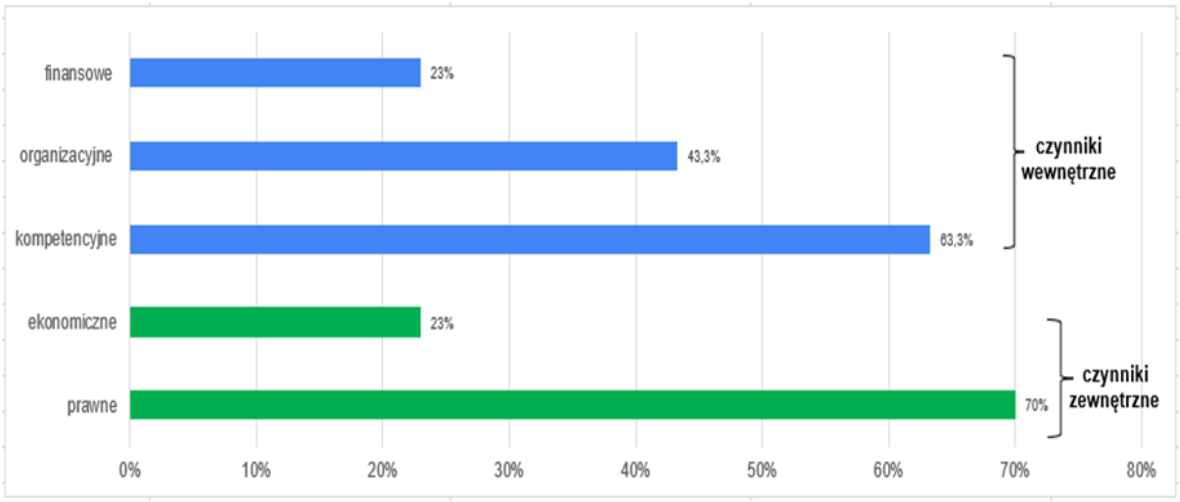
Bariery i ograniczenia spowalniające dynamikę rozwoju otwartych danych w Polsce, w tym w kontekście AI

Poziom obecnego wykorzystania otwartych danych w Polsce, w opinii badanych przedsiębiorstw, okazał się być determinowany zarówno przez czynniki o charakterze wewnętrznym, leżącymi po stronie podmiotów działających obecnie na rynku, jak i zewnętrznym, niezależnych od bieżącej aktywności przedsiębiorstwa. Badane firmy, dokonały wyboru tych czynników, które obecnie mają według nich największe znaczenie w omawianym obszarze. Za największą barierę uznano ograniczenia kompetencyjne zarówno wśród kadry zarządzającej, jak i pracowników, na co wskazało nieco ponad 63% badanych. Przesłanek takiego stanu należy upatrywać zarówno w obecnym poziomie świadomości i wiedzy polskich przedsiębiorstw w tym zakresie, a także ich umiejętności do efektywnego i skutecznego wykorzystania otwartych danych do realizacji stawianych sobie celów biznesowych. W opinii respondentów dotyczy to zarówno *„identyfikacji potrzeb biznesu w tym zakresie, jak i posiadanych możliwości”*, a także *„aspektów mentalnych, zwłaszcza w odniesieniu do kadry zarządzającej”*. Dla 43% badanych przedsiębiorstw, ważnym czynnikiem mogą okazać się także kwestie organizacyjne dotyczące samej struktury podmiotów, zakresu uprawnień decyzyjnych na poszczególnych szczeblach zarządczych, w tym decyzji dotyczących kwestii technologicznych czy zakresu swobody udzielanej w tym zakresie pracownikom. Nieco ponad 23% badanych wskazywało natomiast na aspekty finansowe jako czynnik ograniczający dynamikę rozwoju otwartych danych w Polsce, w tym w kontekście rozwoju sztucznej inteligencji. Będzie miało to przełożenie zarówno na ocenę zasadności ponoszonych w tym zakresie kosztów, jak i ocenę rzeczywistych korzyści, jakie z tego tytułu mogą one osiągnąć, tym samym wpływając na ich skłonność do podjęcia aktywności w analizowanym zakresie.

Wśród czynników zewnętrznych, mających istotny wpływ na analizowane zagadnienia respondenci wyodrębnili dwie grupy o znaczeniu priorytetowym. Przede wszystkim istotna okazała się być grupa czynników prawnych, co nie wydaje się być zaskoczeniem. Wskazało na nią 70% badanych. Uznano, że działania w tym zakresie powinny, z jednej strony prowadzić do systematycznego podnoszenia jakość i ilość udostępnianych danych, stanowiąc tym samym jedną z zachęt do korzystania z nich, z drugiej zaś powinny one w sposób jednoznaczny regulować zasady udostępniania danych w ujęciu sektorowym i/lub cross-sektorowym.

Drugą, istotną grupą czynników okazały się czynniki ekonomiczne, na które wskazało 23% respondentów. W ich opinii mają one między innymi istotny wpływ na decyzje podmiotów rynkowych o poziomie, zakresie i charakterze podejmowanych decyzji inwestycyjnych, w tym zwłaszcza tych dotyczących technologii cyfrowych. Okażą się one także ważne dla szybkości przeprowadzanych w tym zakresie zmian i eksplorację kolejnych obszarów, w których otwarte dane i AI mogą mieć dla nich kluczowe znaczenie w przyszłości.

Wykres 7. *Kluczowe czynniki wewnętrzne i zewnętrzne mające wpływ na dynamikę rozwoju otwartych danych, w tym na rzecz AI, w Polsce, w opinii respondentów*



Źródło: badania własne

Wstępna analiza czynników, które spowalniają dynamikę rozwoju otwartych danych w Polsce, w tym w kontekście dalszego rozwoju sztucznej inteligencji, niewątpliwie wskazuje na konieczność holistycznego i wielowymiarowego spojrzenia na nie. Kompleksowość podejścia, uwzględniająca synergię szerokiego spektrum determinant wewnętrznych i zewnętrznych wraz z ich wzajemnymi zależnościami wydaje się bezsporna, pozwalając na optymalizację realizowanych modeli biznesowych, unikatowość oferowanych na rynku innowacji i zdolności do budowania swojej przewagi konkurencyjnej, w tym ze szczególnym uwzględnieniem sztucznej inteligencji.

Należy pracować nad identyfikacją wszelkich ograniczeń (prawnych, technicznych czy organizacyjnych), które sprawiają, że wymiana danych nie następuje lub napotyka na problemy. Działaniom tym winna zaś towarzyszyć edukacja użytkowników systemów, dysponentów danych, ale i właściwych regulatorów, którzy powinni dobrze zrozumieć nie tylko zasady działania i cele projektów wykorzystujących dane, ale także ich wagę.

Zadaniem władz publicznych powinno być także znoszenie, ciągle jeszcze istniejących, nieuzasadnionych (w tym nieproporcjonalnych lub dyskryminujących) ograniczeń w dostępie do danych, takich jak: koszty (zbyt wygórowane lub nakładane wedle niejasnych kryteriów) lub inne wymogi obciążające wnioskodawców (gdy dostęp do danych wymaga złożenia i rozpatrzenia stosownego wniosku). Powyższa konkluzja nie stanowi postulatu całkowitego zniesienia odpłatności lub innych warunków związanych z dostępem do danych. Czasem ich dokonanie lub spełnienie jest konieczne i zasadne. System dostępu powinien być jednak jasny dla jego użytkowników, a ewentualne ograniczenia – proporcjonalne do rzeczowej potrzeby.

Należy też podkreślić, że systemy współdzielenia danych powinny, poza ściśle określonymi wyjątkami, opierać się na zasadzie dobrowolności. Ze względu na konieczność ochrony takich wartości jak: prywatność czy tajemnice handlowe, z największą ostrożnością należy więc podchodzić do wszelkich koncepcji zakładających przymusowe pozyskiwanie danych od podmiotów prywatnych.

Piotr Zawadzki
Radca prawny i rzecznik patentowy w kancelarii Bird&Bird

Prowadzone badania miały charakter wstępny, a niewielka próba badawcza, na której zostały one zrealizowane, powoduje, że prezentowane wyniki mają przede wszystkim charakter informacyjny i wskazują na ogólne tendencje w omawianym obszarze. Dostarczyły one jednak przesłanek, które stały się podstawą do dalszej eksploracji analizowanych zagadnień oraz krytycznej refleksji do dalszych rozważań w tym zakresie. Pozwoliły także na sformułowanie kilku preliminarnych rekomendacji. Wsłuchując się w opinie przedsiębiorstw, kluczowe jest:

- zintensyfikowanie działań na rzecz budowania świadomości i wiedzy na temat otwartych danych oraz zasad współdzielenia danych, ze szczególnym uwzględnieniem obaw, jakie zgłaszają obecnie, w tym zakresie, przedsiębiorstwa;
- inicjowanie działań ukierunkowanych na współtworzenie przez przedsiębiorstwa rozwiązań korzystnych dla nich z punktu widzenia zarówno korzystania z otwartych danych na rzecz AI, jak i jednoczesnego ich udostępniania;
- wsparcie procesu uwalniania danych między innymi poprzez stworzenie piaskownic pod kątem danych osobowych;
- wzrost efektywności wykorzystania danych pochodzących z repozytoriów, poprzez systematyczne zwiększanie zarówno ilości udostępnianych danych, w tym w czasie rzeczywistym, jak i ich jakości (np. formaty pożądane przez przedsiębiorstwa z punktu widzenia prowadzonej przez nie działalności);
- wyposażenie repozytoriów otwartych danych w narzędzia do ich filtrowania pod kątem możliwości ich wykorzystania do uczenia maszynowego oraz do etykietowania (anotacji danych);
- zacieśnianie współpracy i kooperacji między biznesem a światem naukowo-badawczym (otwarte dane naukowe, otwarte dane badawcze) na rzecz rozwoju AI;
- budowanie zaangażowanych i interdyscyplinarnych społeczności zainteresowanych problematyką otwartych danych i ich znaczeniem w rozwoju sztucznej inteligencji, ze szczególnym uwzględnieniem Polski.

Bibliografia

Borowik M., Maśniak L., Kroplewski R., Romaniec H. (2017) *Przemysł + Gospodarka oparta o dane*, Ministerstwo Cyfryzacji, https://www.gov.pl/web/cyfryzacja/gospodarka-oparta-o-dane-przemysl [dostęp: 28.10.2022].

Brynjolfsson E., McAffe A. (2016), *The second machine age*, John Wiley & Sons, New York-London.

Castrounis A., (2019), *AI for people and business,* O'Reilly Media, Inc, USA.

Gregor B., Kaczorowska-Spychalska D. (2020), *Technologie cyfrowe w biznesie. Przedsiębiorstwa 4.0 a sztuczna inteligencja*, PWN, Warszawa.

Iwasiński Ł. (2016), *Społeczne zagrożenia danetyzacji rzeczywistości*, [w:] Przastek-Samokowa M., Sosińska-Kalata B., Wiorogórska Z. (red.), *Nauka o informacji w okresie zmian. Informatologia i humanistyka cyfrowa*, Wydawnictwo SBP, Warszawa.

Kaplan A., Haenlein M. (2019), *Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence*, „Business Horizons", vol. 62, no. 1.

Kaczorowska-Spychalska D., Sułkowski Ł. (2021), *Determinants of the adoption of AI wearables – practical implications for marketing*, „Human Technology", vol. 17, no 3.

Poniewierski A. (2020), *SPEED. Bez granic w cyfrowym świecie*, www.speednolimits.com, Warszawa.

Romaniuk R. (2020), *Systemy informatyczne jako fundament przedsiębiorstwa 4.0* [w:] B. Gregor, D. Kaczorowska-Spychalska, *Technologie cyfrowe w biznesie. Przedsiębiorstwa 4.0 a sztuczna inteligencja*, PWN, Warszawa.

van Hesteren D., van Knippenberg L. (2021), *Open Data Maturity Report*, Luxembourg: Publications Office of the European Union.

Wawrzyniak B., Musidłowska M., Zygmuntowski J.J. (2022), *Uwolnić potencjał danych.*

Polityka dla rozwoju sztucznej inteligencji w Polsce od roku 2020 (praca zbiorowa) (2020), załącznik do uchwały nr 196 Rady Ministrów z dnia 28 grudnia 2020 r. (poz. 23).

Ustawa z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego (Dz. U. 2021 poz. 1641).

http://opendefinition.org/ [dostęp: 11.08.2022].

https://www.oecd.org/gov/digital-government/open-government-data.htm [dostęp: 17.08.2022].

https://dane.gov.pl/ [dostęp: 18.08.2022].

http://opendefinition.org/ [dostęp: 11.08.2022].

https://opensource.org/licenses/MIT [dostęp: 11.08.2022].

https://eur-lex.europa.eu/legal-content/PL/ALL/?uri=CELEX%3A32019L1024 [dostęp: 23.08.2022].

https://data.gov/ [dostęp: 11.08.2022].

https://archive.ics.uci.edu/ [dostęp: 11.08.2022].

https://dataportals.org/ [dostęp: 11.08.2022].

https://data.europa.eu/ [dostęp: 11.08.2022].

GRAI
GRUPA ROBOCZA
DS. SZTUCZNEJ INTELIGENCJI

CENTRUM INTELIGENTNYCH TECHNOLOGII
Wydział Zarządzania
Uniwersytet Łódzki

DEVINITI